The Model Knows Where to Look: Visual-Grounding-Guided Head Suppression for Mitigating Hallucination in LVLMs

Mong Yuan Sim[1,2] **Wei Emma Zhang**[1] **Xiang Dai**[2] **Byron Fang**[3]
[1]Adelaide University [2]CSIRO [3]Oracle
{mongyuan.sim;wei.e.zhang}@adelaide.edu.au

Abstract

Large Vision-Language Models (LVLMs) achieve strong multimodal reasoning capabilities, yet are still prone to object hallucination - often generating content that is not grounded by the visual input. Existing attention-based training-free mitigating methods rely on visual attention heads identified by high attention weights or sensitivity to visual token perturbations. However, we show that these identified visual attention heads are not necessarily responsible for visual grounding, and using this as guidance is suboptimal. Motivated by this observation, we propose a novel training-free decoding method that identifies visually sensitive heads by counterfactual region masking. We find that these heads provide a clear signal when visual grounding is required and can therefore be used to mitigate hallucination by guiding the suppression of hallucination-prone heads. Across multiple LVLMs on image captioning and VQA benchmarks, our method achieves best or near-best performance in reducing hallucination while preserving overall generation quality. [1]

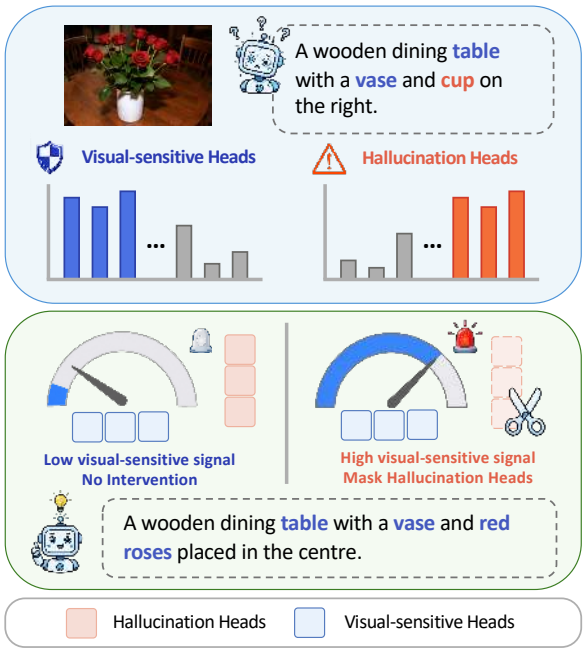


Figure 1: Visual-sensitive heads, identified by our counterfactual region masking method, provide a grounding signal by attending to visual tokens when generation requires image evidence. When this signal exceeds a threshold, we suppress hallucination heads during decoding, reducing unsupported objects while preserving grounded visual content.

1 Introduction

Large vision-language models (LVLMs) have shown impressive performance in open-ended vision-language tasks, such as image captioning, visual question answering (VQA), and multimodal reasoning (Liu et al., 2023, 2024c; Tong et al., 2024; Zhu et al., 2024). However, they are still prone to object hallucination, where generated responses mention objects, attributes, or relations not grounded in the input image. Among the many explanations for this phenomenon (Yang et al., 2025; Liu et al., 2024b,a; Hu et al., 2023), a widely accepted view is that LVLMs often fail to effectively utilize visual information due to their over-reliance on language priors, with their internal representations encoding strong textual regularities and co-occurrence patterns (Liu et al., 2024d; Huang et al., 2024; Sim et al., 2025; Viltoriano et al., 2026).

Correspondingly, a promising class of training-free methods mitigates object hallucination through direct attention head-level intervention, e.g., by amplifying visually aware heads or suppressing hallucination heads (Sarkar et al., 2025; He et al., 2025; Yang et al., 2025; Yin et al., 2025). Despite encouraging results, existing attention-based methods still face two key limitations.

First, existing methods identify visual-related heads using visual attention weights directly (Liu

[1]Our code is available at: https://github.com/raymondsim/grounding-hallucination-heads.

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 3347–3367
November 6-10, 2026 ©2026 Association for Computational Linguistics

et al., 2024d; Sarkar et al., 2025) or attention changes by removing images (He et al., 2025). High attention weights on image tokens do not necessarily indicate that a head uses relevant visual evidence, making attention alone an unreliable signal for identifying visually sensitive heads. This limitation is closely related to the long-standing debate on whether attention weights can be directly interpreted as reliable indicators of input importance (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019). To address this limitation, we propose to use counterfactual region-masking to identify visual-sensitive heads, which play a more direct role in grounding visual evidence in image tokens.

In addition, existing head-level interventions often apply suppression or enhancement in a relatively uniform manner. They suppress hallucination-prone heads or reinforce visual-related heads throughout generation, regardless of whether the current decoding step requires visual evidence. Such unconditional intervention can lead to shorter descriptions, lower recall, or degrade answer quality. To resolve this, we use the identified visual-sensitive heads as a step-level signal of visual demand. During decoding, the signal determines when grounding is necessary and hallucination heads should be suppressed, enabling a more targeted intervention only when stronger visual grounding is needed.

Our contributions are summarized as follows:

- We propose a counterfactual region-masking method to identify visual-sensitive heads as a distinct set of attention heads that are more directly involved in grounding visual evidence within image tokens.
- We show that these heads provide a reliable internal signal for visual demand, revealing when generation requires image evidence rather than language priors.
- We introduce a training-free decoding intervention that uses signal to conditionally suppress hallucination heads, reducing hallucination while preserving generation quality. We demonstrate its effectiveness across multiple benchmarks for image captioning and VQA.

2 Related Work

Hallucination Mitigation in LVLMs. Approaches to mitigating hallucination can be divided into building better training data (Hu et al., 2023; Liu et al., 2024a) and training-free methods. We focus on the latter, which can be directly applied to existing LVLMs without additional training. Early work, such as VCD (Leng et al., 2024) contrasts the logits distribution between the original and distorted visual inputs to reduce LVLMs' over-reliance on the language prior, and later variants use image-biased predictions (Zhu et al., 2025), and self-generated visual descriptions (Kim et al., 2024) as reference distributions. Other methods introduce external or structured visual guidance, including OPERA's summary token penalty and rollback strategy (Huang et al., 2024), HALC's fine-grained grounding-based correction (Chen et al., 2024), and CLIP-guided decoding (Deng et al., 2024). Most related to our work, SPIN (Sarkar et al., 2025) selects top-k heads with the highest attention to visual tokens as visual-sensitive heads, and suppresses the rest of the heads, while VHR (He et al., 2025) enhances visually sensitive heads, ranked by attention divergence with and without visual tokens. In contrast, our work uses visual-sensitive heads as a step-level signal to decide when hallucination heads should be suppressed.

Head-Level Multimodal Interpretability. Recent head-level analyses suggest that LVLM behavior is often concentrated in a small subset of attention heads. These heads can specialize in functions such as entity and syntactic grounding (Li et al., 2020), visual-token processing (Bi et al., 2025), multimodal reasoning (Jiang et al., 2025), controllable semantic or visual concepts (Basile et al., 2025). Kang et al. (2025b) show that only a few attention heads in frozen LVLMs can provide strong visual grounding signals, and use these localization heads for training-free object grounding. From the hallucination perspective, Yang et al. (2025) show through causal attribution and intervention that multi-head attention modules strongly contribute to hallucinated tokens, while Kang et al. (2025a) identify visual attention sinks, where LVLMs assign high attention to visually irrelevant tokens. In line with these studies, our work focuses on identifying functionally meaningful attention heads, using counterfactual region masking to capture heads whose behavior is directly affected by grounded visual evidence.

3 Preliminary

A typical LVLM consists of an image encoder, a pretrained Large Language Model (LLM), and a

connector that maps visual representations to the LLM's representation space, where they are processed as visual tokens by the LLM. The language backbone of modern LVLMs is typically built upon Transformer layers with multi-head self-attention (MHA). Given input features $X_l \in \mathbb{R}^{N \times d_{\text{model}}}$ at layer l, where N denotes the total number of input tokens, MHA first projects X_l into query, key, and value matrices for each attention head. Then, the head output of each head is computed using scaled dot-product attention:

$$A_{l,h} = \text{softmax}\left(\frac{Q_{l,h}K_{l,h}^{\top}}{\sqrt{d_h}}\right), \quad O_{l,h} = A_{l,h}V_{l,h}, \tag{1}$$

where $Q_{l,h}$, $K_{l,h}$, and $V_{l,h}$ are the head-specific query, key, and value representations, d_h is the head dimension, and $A_{l,h} \in \mathbb{R}^{N \times N}$ denotes the attention matrix of head h at layer l. Through this mechanism, different heads can attend to different parts of the input tokens.

The multi-head output is finally obtained by concatenating all head outputs and projecting them back into the model hidden space:

$$\text{MHA}(X_l) = \text{Concat}(O_{l,1}, \ldots, O_{l,H})W_l^O. \tag{2}$$

In LVLMs, the input tokens consist of visual tokens, textual prompt tokens, and previously generated tokens. We denote the index set of visual tokens as $\mathcal{V} \subseteq \{1, \ldots, N\}$. This allows us to quantify how much attention a given head assigns to the visual context by aggregating $A_{l,h}^{i,j}$ over $j \in \mathcal{V}$.

4 Methodology

4.1 Detecting Functional Heads

As discussed earlier, an effective hallucination mitigation method requires deciding both *when* to intervene and *where* to intervene, we consider two complementary types of functional heads. First, *visual-sensitive heads* respond to grounded visual evidence and serve as a step-level signal for when visual grounding is required. Second, *hallucination heads* are associated with unsupported visual generation and define the targets for intervention.

Visual Sensitive Heads. For visual-sensitive heads, our goal is to identify heads that respond to changes in the visual evidence supporting the generated content. Previous work mainly identifies visually important heads by selecting the top-k attention heads that assign high attention weights to visual tokens (Sarkar et al., 2025; Kang et al., 2025b). However, high attention weights to visual tokens does not necessarily imply meaningful visual grounding, as some heads may attend heavily to irrelevant visual tokens, a phenomenon known as the visual attention sink (Kang et al., 2025a). Moreover, the extent to which attention weights reflect information or content importance remains a long-standing debate in the literature (Wiegreffe and Pinter, 2019; Jain and Wallace, 2019). Therefore, we identify visual-sensitive heads using region-level masking ablation, which measures whether a head responds to changes in grounded visual evidence.

Specifically, we randomly sample 500 examples from a grounding dataset, RefCOCO+ (Yu et al., 2016), where each example consists of a natural language phrase describing a target object. For each example, we run inference twice: once using the original image, and once using a counterfactual image where the grounded object region is masked. We then compare the head-level attention differences to identify visual-sensitive heads. Intuitively, this score measures how much the attention pattern of a head changes when the visual evidence for the grounded object is removed. A larger value indicates that the head is more sensitive to the visual region relevant to the object.

Formally, for each example x, let $\tilde{x}$ denote its counterfactual version where the grounded object region is masked. We define the visual sensitivity score of head h in layer l as:

$$S^{(l,h)}(x) = \frac{1}{|\mathcal{Q}_B||\mathcal{V}_B|} \sum_{i \in \mathcal{Q}_B} \sum_{j \in \mathcal{V}_B} \left| A_{l,h}^{i,j}(x) - A_{l,h}^{i,j}(\tilde{x}) \right|. \tag{3}$$

Here, $\mathcal{Q}_B$ denotes the set of query-token positions associated with the target phrase or generated object B, and $\mathcal{V}_B$ denotes the set of visual-token positions corresponding to the masked grounded region. Following the notation in Section 3, $A_{l,h}^{i,j}$ is the attention weight assigned by query token i to key token j in head h of layer l.

To obtain a dataset-level visual sensitivity score, we average the sample-level scores over all examples in $\mathcal{D}$:

$$\bar{S}^{(l,h)} = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} S^{(l,h)}(x). \tag{4}$$

We then rank attention heads according to $\bar{S}^{(l,h)}$, where heads with higher scores are considered

more visually sensitive. These heads are expected to contribute more strongly to visual grounding because their attention distributions are substantially affected by perturbations to the corresponding visual evidence.

Hallucination Heads. To detect hallucination-related heads, we first run inference on 1,000 randomly sampled images from MSCOCO using greedy decoding. We then apply the CHAIR evaluation to identify hallucinated object tokens in each generated caption. Given an input image x, prompt q, and previously generated tokens $y_{<t}$, let y_t denote a hallucinated object token generated at decoding step t. We identify hallucination-related heads by measuring how much each attention head contributes to the model's confidence in generating y_t. Formally, we define the contribution score of head (l, h) as the decrease in the log-probability of y_t when this head is masked:

$$\begin{aligned}\Delta_t^{(l,h)} = &\log p_\theta(y_t \mid x, q, y_{<t}) \\ &- \log p_{\theta \setminus (l,h)}(y_t \mid x, q, y_{<t}),\end{aligned} \tag{5}$$

where p_θ denotes the original model distribution, and $p_{\theta \setminus (l,h)}$ denotes the model distribution after masking head (l, h). Following the notation in Section 3, masking a head means removing its head output $O_{l,h}$ from the multi-head attention computation. A larger positive value of $\Delta_t^{(l,h)}$ indicates that head (l, h) increases the likelihood of the hallucinated token, suggesting a stronger contribution to hallucination.

For each hallucinated token occurrence $i \in \mathcal{E}$, we compute the corresponding score $\Delta_i^{(l,h)}$. The final hallucination score for head (l, h) is obtained by averaging over all hallucinated token occurrences:

$$S_{\text{hal}}^{(l,h)} = \frac{1}{|\mathcal{E}|} \sum_{i \in \mathcal{E}} \Delta_i^{(l,h)}. \tag{6}$$

Here, $\mathcal{E}$ denotes the set of hallucinated token events collected from the dataset. We rank attention heads according to $S_{\text{hal}}^{(l,h)}$, where heads with larger scores are considered more hallucination-inducing because masking them causes a larger reduction in the probability assigned to hallucinated tokens.

4.2 Visual-Gated Hallucination Head Suppression

The head detection procedure described above yields two head sets: visual-sensitive heads, which track visual grounding signals, and hallucination heads, which are associated with unsupported visual generation We use these two head sets to build a training-free decoding intervention method that leverages visually sensitive attention heads as an internal signal for visual-grounded generation. Our key intuition is that hallucination mitigation should not be applied uniformly across all decoding steps, or solely by comparing attention mass between text and visual tokens. Instead of uniformly modifying all decoding steps and blindly relying on the attention mass given to either text or visual tokens, we use the attention behaviour of visually sensitive heads as a step-level grounding signal. Therefore, we selectively intervene at such steps by suppressing hallucination-inducing heads and strengthening the contribution of visually sensitive heads, encouraging the model to rely more on visual evidence when generating visually grounded tokens.

Let $x_{1:T}$ denote the multimodal input sequence, and let the index set of visual tokens be $\mathcal{V} \subseteq \{1, \ldots, T\}$. For Transformer layer l and attention head h, let $A^{(l,h)} \in \mathbb{R}^{T \times T}$ be the attention matrix, where $A_{t,j}^{(l,h)}$ is the attention weight assigned at decoding step t to token j. The output of head h at step t is

$$o_t^{(l,h)} = \sum_{j=1}^{T} A_{t,j}^{(l,h)} W_V^{(l,h)} x_j. \tag{7}$$

Let $\mathcal{H}_{\text{hal}}$ denote the set of hallucination heads to be suppressed, and let $\mathcal{H}_{\text{vis}}$ denote the set of visual-sensitive heads used for gating. For a visual-sensitive head $v \in \mathcal{H}_{\text{vis}}$, define its visual attention mass at decoding step t as

$$s_t^{(v)} = \sum_{j \in \mathcal{V}} A_{t,j}^{(v)}. \tag{8}$$

When multiple visual-sensitive heads are used, we aggregate them as

$$s_t = \max_{v \in \mathcal{H}_{\text{vis}}} s_t^{(v)}. \tag{9}$$

For the single-head case, this reduces to $s_t = s_t^{(v)}$.

The proposed intervention softly suppresses each hallucination head $(l, h) \in \mathcal{H}_{\text{hal}}$ whenever the visual attention mass exceeds the threshold τ:

$$\tilde{o}_t^{(l,h)} = \begin{cases} \alpha \cdot o_t^{(l,h)}, & \text{if } s_t > \tau, \\ o_t^{(l,h)}, & \text{otherwise}, \end{cases} \quad (l, h) \in \mathcal{H}_{\text{hal}}, \tag{10}$$

where $\alpha \in [0, 1]$ controls the suppression strength. When $\alpha = 0$, the intervention reduces to hard masking; when $\alpha = 1$, no suppression is applied.

In words, at each decoding step, we measure how much the designated visual-sensitive heads attend to visual tokens. If this visual attention mass exceeds a threshold, we completely deactivate the selected hallucination heads by setting their outputs to zero for that step.

5 Experiments

LVLMs. We conduct experiments on representative LVLMs with different training strategies and backbone models, i.e., LLaVA-1.5-7b (Liu et al., 2023), Qwen3-VL-8B-Instruct (Team, 2025), and DeepSeek-VL2-Small (Wu et al., 2024).

Baselines. We compare our proposed method with recent baselines: VCD (Leng et al., 2024) uses contrastive decoding to contrast the logits distribution between original and distorted images; OPERA (Huang et al., 2024) mitigates the over-trust issue by using a penalty term in beam-search decoding; PAI (Liu et al., 2024d) amplify attention to visual tokens and subtract logits with no visual input; DAMRO (Gong et al., 2024) uses contrastive decoding to reduce influence of noisy visual tokens; ADHH (Yang et al., 2025) identify hallucination heads and suppress condition on text attention weights; SPIN (Sarkar et al., 2025) assumes top-k attention head with high attention weights on visual tokens as visually important heads, and suppresses the rest of the heads; VHR (He et al., 2025) mitigates hallucination by enhancing vision-aware attention heads at decoding.

Datasets. We evaluate our proposed method and baselines on image captioning and VQA tasks curated to test models' performance on vision-centric questions. For image captioning, we use MSCOCO validation set (Lin et al., 2014). Following previous work (Yang et al., 2025; Sarkar et al., 2025; He et al., 2025), we randomly sample 1000 images from each dataset and generate image captions using LVLMs. For vision-centric VQA, we use POPE (Li et al., 2023) and MME (Fu et al., 2023). POPE is a dataset for object hallucination evaluation. It consists of 500 images from MSCOCO and has three splits: random, popular, and adversarial. MME contains 14 sub-tasks, covering the perception and cognitive abilities of LVLMs. MME uses accuracy and accuracy+ scores as its metric, we explain the score calculation in Appendix A. We also evaluate the performance of our method on LLaVA-Bench (In-the-Wild) (Liu et al., 2024c) which includes a wide range of vision-centric challenging tasks. The outputs from LVLMs on LLaVA-Bench are evaluated using GPT-5-mini, the prompt is provided in Appendix F.

Evaluation Metrics. To evaluate the performance of our methods in VQA tasks, we use F1 and Accuracy scores for POPE. For MME, we use the official Accuracy and Accuracy$^+$ metrics (see Appendix A). The outputs from LVLMs on LLaVA-Bench are evaluated using GPT-5-mini, the prompt is provided in Appendix F. To evaluate object hallucination in image captioning, we use Caption Hallucination Assessment with Image Relevance (CHAIR) (Rohrbach et al., 2018) It has two variants: CHAIR_i, which calculates the fraction of object instances that are hallucinated, and CHAIR_s, which calculates the fraction of sentences containing at least one hallucinated object. Specifically,

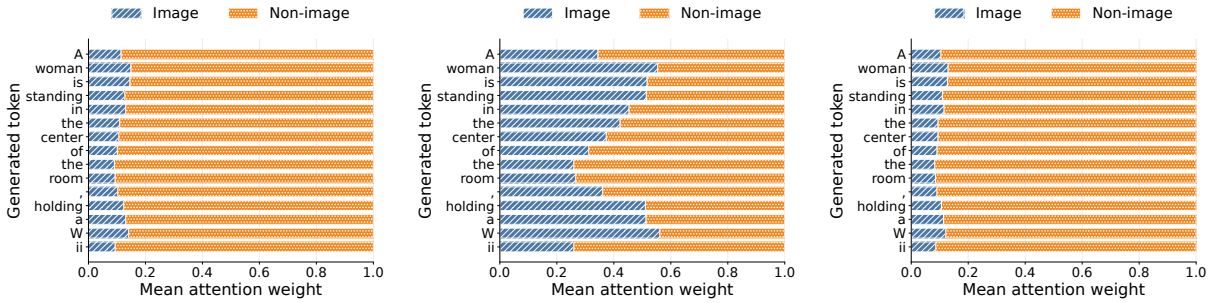
$$\text{CHAIR}_S = \frac{|\{\text{caption w/ hallucinated objects}\}|}{|\{\text{all captions}\}|},$$
$$\text{CHAIR}_I = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all mentioned objects}\}|}. \quad (11)$$

Implementation Details. For our proposed method, we set α and τ in Equation (10) to 0.0 and 0.3, respectively and $k = 50$ for visual-sensitive heads and hallucination heads. We provide an extensive ablation study on these hyperparameters in Appendix C. All hyperparameters for baseline methods are set to default, based on the official implementation, details are provided in Appendix D. All experiments were conducted on a single NVIDIA H100 94GB GPU.

5.1 Visual-Sensitive Heads Analysis

	LLaVA-1.5		Qwen3-VL-8B		Deepseek2-VL	
# Masked Heads	20	100	20	100	20	100
Baseline	0.6901		0.8800		0.770	
Random Heads	0.5662	0.3680	0.8600	0.7743	0.7583	0.7703
Random Mask	0.7470	0.6238	0.8621	0.8420	0.7811	0.7680
VHR	0.5662	0.5598	0.8800	0.8800	0.7972	0.6877
SPIN	0.5867	0.3552	0.8601	0.6875	0.7994	0.7200
Ours	**0.3100**	**0.2800**	**0.7500**	**0.6050**	**0.6458**	**0.5238**

Table 1: Performance on MME dataset, masking attention heads using different methods, evaluated using F1 score. Lower performance indicates that the masked heads are more important for the target behavior, as removing them causes a larger degradation.



(a) Hallucination Heads (b) Visual-sensitive Heads (c) Other Heads

Figure 2: Attention weights distribution on image and non-image tokens for hallucination heads, visual-sensitive heads, and others. Visual-sensitive heads provide high attention weights to image tokens compared to hallucination and other heads. The attention weights peak is often at object tokens like "woman" and "Wii", showing the tendency of visual-sensitive heads giving high attention weights when the model requires visual grounding capability.

Causal Effect of Visual Sensitive Heads on VQA We assess whether the detected visually sensitive heads are functionally important by masking the top-k heads selected using different methods and evaluating the resulting performance on 100 randomly sampled VQA samples from the MME dataset (Fu et al., 2023). For SPIN and VHR, we dynamically mask top-k heads selected for each sample, following their original selection strategy. In contrast, our method uses a fixed set of top-k heads identified offline and shared across the entire dataset. As shown in Table 1, masking our detected visual-sensitive heads leads to a substantial performance drop, indicating that these heads are important for vision-centric tasks. Compared with existing head-selection methods, our detected heads cause a larger degradation when masked under the same masking budget, suggesting that our ranking method better identifies attention heads that are more important for visual grounding.

Detected Heads Are Consistent Across Samples To examine whether the ranked head groups are stable across data samples, we perform a consistency analysis. For visual-sensitive and hallucination heads, we randomly sample non-overlapping 500 and 1000 samples, respectively, independently compute head rankings on each subset, and compare the resulting top-k sets using Overlap@k and Jaccard@k.

Figure 3 shows the consistency of ranked head groups across different random subsets of the data. Visual-sensitive heads exhibit remarkably high stability, achieving near-perfect Overlap@k and Jaccard@k across all values of k. In contrast, hallucination heads show moderate consistency across subsets, where Overlap@k and Jaccard@k are above 0.5 when k is greater than 50, suggesting that a reasonably stable subset of hallucination-related heads emerges as more heads are considered. As a comparison, we also compare the consistency between two random sets of heads, which shows extremely low consistency compared to the ranked visual-sensitive and hallucination heads.

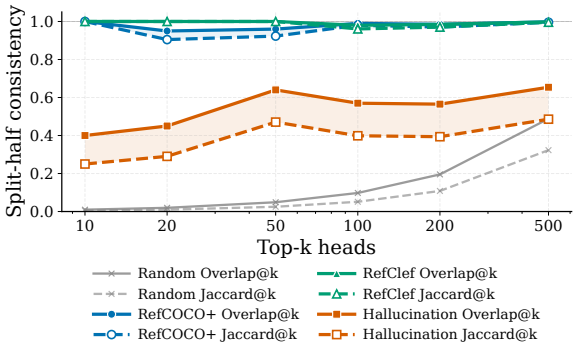


Figure 3: Visual-sensitive and hallucination heads consistency against random heads. Visual-sensitive heads show near-perfect agreement, while hallucination heads show moderate consistency, with Overlap@k and Jaccard@k above 0.5 when k is greater than 50.

To further test whether this stability is dataset-specific, we repeat the visual-sensitive head ranking procedure on RefClef (Rohrbach et al., 2016). Specifically, we randomly sample two non-overlapping subsets of 500 examples, apply the same region-masking procedure, and compute head rankings independently. We observe that visual-sensitive heads identified on RefClef achieve near-perfect Overlap@k and Jaccard@k across different values of k. This cross-dataset result supports that our counterfactual region-masking criterion captures stable evidence-sensitive heads, rather than heads specific to a particular dataset split.

	LLaVA-1.5-7B			Qwen3-VL-8B			Deepseek2-VL		
	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	F1 ↑	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	F1 ↑	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	F1 ↑
Greedy	46.5	12.9	76.7	56.1	10.5	74.3	50.5	14.7	72.6
VCD	53.7	15.9	74.7	58.0	10.9	74.8	50.7	15.0	72.9
OPERA	46.4	12.9	76.7	53.0	9.6	75.0	-	-	-
PAI	33.0	10.2	74.5	55.4	10.5	75.0	51.3	14.4	**73.1**
DAMRO	45.8	12.7	**76.9**	51.9	9.9	**75.4**	50.8	14.8	72.6
ADHH	26.2	8.4	74.1	50.5	8.8	73.6	42.5	18.1	71.1
SPIN	25.6	7.19	73.5	-	-	-	43.4	14.8	71.7
VHR	31.5	10.2	74.7	-	-	-	43.4	14.7	72.6
Ours	**25.0**	**7.8**	74.2	**45.6**	**7.4**	73.6	**41.7**	**14.3**	71.4

Table 2: The performance of different decoding intervention methods on the MSCOCO validation set. * Results are not reported where the recommended hyperparameters produced degenerate or otherwise unreliable generation behavior in our setup.

Method	LLaVA	Qwen3-VL	Deepseek-VL
Greedy	85.66	87.89	87.40
VCD	85.05	88.44	**89.34**
OPERA	85.51	87.95	87.36
PAI	67.04*	87.86	88.21
DAMRO	85.51	88.01	87.62
ADHH	67.10*	87.07	88.82
SPIN	66.65*	87.90	89.02
VHR	84.36	87.96	87.40
Ours	**86.18**	**88.70**	89.02

Table 3: Average F1 performance of decoding intervention methods across POPE random, popular, and adversarial settings. Results marked with * are not directly comparable as a substantial proportion of generated responses cannot be parsed as valid yes/no answers.

Model	Method	Percep.	Cog.	Total	Δ
LLaVA-1.5	Baseline	1476.8	**310.0**	1786.8	–
	Ours	**1486.0**	309.5	**1795.5**	+8.7
Qwen3-VL	Baseline	**1713.0**	665.0	2378.0	–
	Ours	1710.8	**670.0**	**2380.8**	+2.8
Deepseek-VL	Baseline	1665.2	**501.4**	2166.6	–
	Ours	**1672.5**	498.2	**2170.7**	+4.1

Table 4: Aggregate MME results across all evaluated LVLMs. We report perception, cognition, and overall scores, with the complete category-wise results provided in Appendix B.

Visual Sensitive Heads Signal When Visual Grounding Is Needed. After categorizing attention heads into visually sensitive and hallucination-inducing groups, we further examine how these heads allocate attention to image and text tokens during each generation step. As shown in Figure 2, visually sensitive heads exhibit a stronger tendency to attend to visual tokens when the model generates visually grounded content. In particular, attention to image tokens increases around object-related tokens, such as "Wii", "woman", and "group", as well as tokens immediately following these object mentions. Meanwhile, attention to function words such as "A" and "the" are lower. This suggests that visually sensitive heads can serve as reliable internal indicators of when visual grounding is required at each generation step.

5.2 Downstream Evaluation

Image Captioning. Table 2 presents the performance of our proposed method, compared to existing intervention methods. Compared with greedy decoding, our method substantially reduces hallucination, lowering $CHAIR_S$ from 46.5 to 25.0 and $CHAIR_I$ from 12.9 to 7.8, corresponding to reductions of 46.2% in sentence-level hallucination and 39.5% in instance-level hallucination. Notably, our method achieves the lowest CHAIR scores among all compared intervention methods, suggesting that grounding-aware head suppression is particularly effective in reducing unsupported object mentions in open-ended generation.

We observe a mild decrease in object-level F1 score compared with VCD and OPERA. This indicates that our intervention makes the model more conservative, meaning that it reduces unsupported objects generation, but may also suppress some correct object mentions. This phenomenon is consistent with prior hallucination mitigation studies, which have observed a trade-off between reducing hallucinated content and preserving caption coverage (Sarkar et al., 2025; He et al., 2025). Nevertheless, our method lies at a balanced point in this trade-off, where it achieves the lowest CHAIR scores and a competitive F1 score. We provide qualitative analyses of representative success and failure cases in Appendix E.

Model	Method	$CHAIR_S$	$CHAIR_I$	Recall	Precision	F1	Len
LLaVA-1.5-7B	VHR	31.5	10.2	**68.5**	82.0	**74.7**	75.0
	SPIN	29.2	8.7	66.4	84.0	74.2	73.2
	Ours	**24.7**	**8.1**	64.3	**87.5**	74.6	76.3
Qwen3-VL-8B	VHR	50.9	9.5	74.3	74.8	74.5	382.7
	SPIN	51.6	9.9	**74.4**	74.8	**74.6**	382.7
	Ours	**45.6**	**7.4**	69.5	**78.1**	73.6	434.8
DeepSeek-VL2-S	VHR	50.5	14.7	**69.8**	75.7	72.6	185.4
	SPIN	45.9	13.4	68.2	76.6	72.2	192.9
	Ours	**42.9**	**11.8**	67.7	**79.1**	**73.0**	178.0

Table 5: Comparison of head selection strategies for attention-based gating. Using the same gating mechanism, our counterfactually selected heads achieve lower CHAIR scores and higher precision across all three LVLMs, with comparable F1 score and a modest reduction in recall.

	Visual Correctness ↑	Relevance ↑	Detailedness ↑
LLaVA	5.12	6.88	**5.10**
+ Ours	**5.53**	**6.91**	4.97
Qwen	5.42	8.50	**6.42**
+ Ours	**5.55**	**8.60**	6.40
Deepseek	6.60	8.45	**6.73**
+ Ours	**7.00**	**8.85**	5.98

Table 6: LLaVA-Bench (In-the-Wild) evaluation results, scored by GPT-5-mini, ranged between 0 to 10.

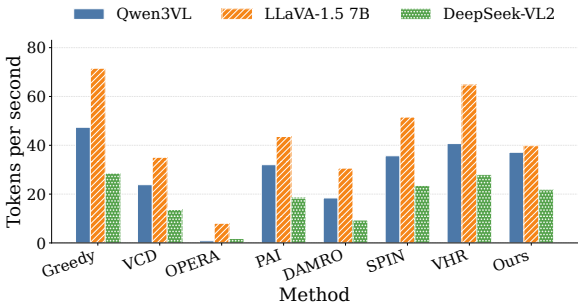


Figure 4: Inference efficiency comparison (tokens per second) across decoding methods. Higher values indicate faster decoding speed.

POPE Evaluation. Polling-based Object Probing Evaluation (POPE) (Li et al., 2023) is a dataset that evaluates hallucination in VQA form by querying models "*Is there* `[object]` *in the image?*". Objects are categorized into random, popular, and adversarial splits. This provides a complementary evaluation to image captioning by testing hallucination under direct object-level questioning. Table 3 summarizes the averaged F1 scores across the random, popular, and adversarial POPE settings. Our method achieves the strongest average F1 on LLaVA and Qwen3-VL, while remaining competitive on Deepseek-VL.

MME Evaluation. MME (Fu et al., 2023) evaluates model's perception and cognition abilities across 14 subtasks, including commonsense reasoning, celebrity, position, and OCR. Table 4 reports the aggregate MME results, and the complete category-wise scores are provided in Appendix B. Across all three LVLMs, our method maintains the baseline cognition and perception performance, and achieves a slight overall performance. These results indicate that our method reduces hallucination without compromising the models' perception and cognition abilities.

LLaVA-Bench (In the Wild) Evaluation. LLaVA-Bench (In-the-Wild) (Liu et al., 2024c) is a challenging benchmark that contains 24 images and 60 questions, crafted to assess model's ability in simple question answering, detailed descriptions, and complex reasoning. The generated outputs are then judged by GPT-5-mini, using the prompt in Appendix F. The evaluation results on LLaVA-Bench (In-the-Wild) are presented in Table 6. These results demonstrate that our method improves LVLMs accuracy on visual correctness and the generated output's relevance to the query. Importantly, these gains are achieved while largely preserving the detailedness of the generated outputs, suggesting that our method mitigates hallucination without simply producing shorter or less informative responses.

5.3 Running Time Analysis

Figure 4 compares inference efficiency in tokens generated per second across different decoding methods. Greedy decoding is the fastest baseline as it performs single-pass inference without any additional intervention. Our method achieves substantially higher decoding speed across models compared to VCD and OPERA, which require

Model	Method	$CHAIR_S$	$CHAIR_I$	Recall	Precision	F1	Len
LLaVA-1.5-13B	–	44.2	11.7	**77.8**	77.9	**77.8**	98.9
	Ours	**30.5**	**9.7**	68.7	**82.8**	75.0	97.3
Qwen-VL3-32B	–	58.0	12.6	**77.0**	72.3	74.6	330.1
	Ours	**54.1**	**11.2**	75.7	**73.8**	**74.7**	331.9
DeepSeek-VL2 (27B)	–	51.0	14.8	**69.6**	**75.2**	**72.3**	188.6
	Ours	**46.9**	**13.5**	67.6	**75.2**	71.2	205.9

Table 7: Image captioning results on larger LVLMs (13-32B), comparing greedy decoding without intervention and our proposed method. Our method consistently reduces CHAIR hallucination across LLaVA-1.5-13B, Qwen3-VL-32B, and Deepseek-VL2 (27B), while exhibiting model-dependent precision-recall trade-offs.

additional passes or expensive beam-search operations during generation. Similar to SPIN and VHR, our method avoids repeated forward passes, with only a small overhead from token-level gating. Together with the hallucination results, this shows our method balances hallucination reduction and decoding efficiency.

5.4 Analysis of the Visual-Sensitive Head Signal

We analyze how the proposed visual gate controls hallucination-head suppression, with full results provided in Appendix C. The visual threshold determines when suppression is activated: lower thresholds apply the intervention more frequently, while higher thresholds require stronger visual-head evidence and therefore suppress more selectively. Across LLaVA, Qwen3-VL, and DeepSeek-VL2, we observe that overly aggressive gating can reduce CHAIR scores but may also suppress grounded object mentions, lowering recall and F1. Conversely, overly conservative gating under-activates the intervention and allows hallucination to return. These results suggest that the visual gate is not merely a tunable strength parameter, but a signal-based mechanism for calibrating when hallucination-head suppression should be applied.

5.5 Comparison of Head Selection Strategies

To examine whether counterfactual head selection method truly identifies visual-sensitive heads that are more causally important than those selected by existing approaches, we apply the same attention aggregation and suppression gate to head sets identified by VHR, SPIN, and our counterfactual method. This controlled comparison isolates the effect of the head selection strategy while keeping the downstream intervention mechanism unchanged. The results are presented in Table 5. Across three LVLMs, our signal consistently yields lower CHAIR scores and higher precision, while maintaining comparable F1 scores with a modest reduction in recall. These results suggest that counterfactual selection identifies heads that provide a more reliable signal of visual grounding, leading to more effective suppression of hallucinated objects during generation.

5.6 Scalability Across Model Sizes

We further conducted image captioning experiments using greedy decoding on larger variants of the evaluated LVLM families, comparing the original models without intervention against our proposed method. Specifically, we evaluated on LLaVA-1.5-13b, Qwen3VL-32B, and 27b variant of Deepseek-VL2 base. Experimental results in Table 7 show that the effectiveness of our method persists at substantially larger model scales. These results suggest that the identified attention-head behaviors and the resulting intervention generalize beyond smaller models, remaining effective in LVLMs ranging from 13B to 32B parameters.

6 Conclusion

In this work, we propose counterfactual region masking to identify visual-sensitive heads responsible for evidence-grounded generation. These heads provide a more reliable grounding signal than visual attention mass and attention-divergence-based head selection. We utilize this signal to suppress hallucination-prone heads, reducing unsupported visual content without affecting generation quality. Across LVLMs and benchmarks, our method consistently reduces hallucination while preserving general multimodal performance. Overall, targeted visual-sensitive head intervention provides a reliable and effective method to reduce hallucination in LVLMs. Future work could extend targeted head interventions to specialized visual capabilities, such as OCR, counting, and spatial reasoning.

Limitations

Despite the strong empirical results of our proposed method, we acknowledge several limitations of our proposed method, and the results should be interpreted with caution:

- Attention ranking and attribution in this paper are performed using general domain datasets (i.e., RefCOCO+ and MSCOCO). As a result, we do not expect the identified hallucination and visual-sensitive heads fully generalize to specialized domains such as the medical domain. Applying our method to VLMs in such domains requires domain-specific head reranking or calibration. Moreover, our visual-sensitive head detection relies on counterfactual region masking, which requires access to image regions that are equivalent to the text caption. While this is feasible for general-domain datasets with grounding annotations, such annotations may be unavailable or expensive to obtain in other specialized domains.

- Our method is primarily evaluated on object hallucination and visual grounding-oriented benchmarks, where hallucination is closely related to unsupported visual content. Although we also evaluate on broader LVLM benchmarks such as LLaVA-Bench (In-the-Wild) and MME, the intervention is mainly designed to improve visual faithfulness during generation. Its effect on tasks requiring complex reasoning, multi-step instruction following, or external knowledge may differ, since such tasks involve additional factors beyond visual grounding.

Acknowledgment

Mong Yuan Sim is supported by CSIRO's Data61 Top-up scholarship. Wei Emma Zhang is supported by the Australian Research Council Early Career Industry Fellowships (IE230100119).

References

Lorenzo Basile, Valentino Maiorca, Diego Doimo, Francesco Locatello, and Alberto Cazzaniga. 2025. Head pursuit: Probing attention specialization in multimodal transformers. *CoRR*, abs/2510.21518.

Jing Bi, Junjia Guo, Yunlong Tang, Lianggong Bruce Wen, Zhang Liu, Bingjie Wang, and Chenliang Xu. 2025. Unveiling visual perception in language models: An attention head analysis approach. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 4135–4144. Computer Vision Foundation / IEEE.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. 2024. HALC: object hallucination reduction via adaptive focal-contrast decoding. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 7824–7846. PMLR / OpenReview.net.

Ailin Deng, Zhirui Chen, and Bryan Hooi. 2024. Seeing is believing: Mitigating hallucination in large vision-language models via clip-guided decoding. *CoRR*, abs/2402.15300.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. MME: A comprehensive evaluation benchmark for multimodal large language models. *CoRR*, abs/2306.13394.

Xuan Gong, Tianshi Ming, Xinpeng Wang, and Zhihua Wei. 2024. DAMRO: dive into the attention mechanism of LVLM to reduce object hallucination. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 7696–7712. Association for Computational Linguistics.

Jinghan He, Kuan Zhu, Haiyun Guo, Junfeng Fang, Zhenglin Hua, Yuheng Jia, Ming Tang, Tat-Seng Chua, and Jinqiao Wang. 2025. Cracking the code of hallucination in lvlms with vision-aware head divergence. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 3488–3501. Association for Computational Linguistics.

Hongyu Hu, Jiyuan Zhang, Minyi Zhao, and Zhenbang Sun. 2023. CIEM: contrastive instruction evaluation method for better instruction tuning. *CoRR*, abs/2309.02301.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. OPERA: alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13418–13427. IEEE.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short*

Papers), pages 3543–3556. Association for Computational Linguistics.

Yanbei Jiang, Xueqi Ma, Shu Liu, Sarah Monazam Erfani, Tongliang Liu, James Bailey, Jey Han Lau, and Krista A. Ehinger. 2025. Investigating the functional roles of attention heads in vision language models: Evidence for reasoning modules. *CoRR*, abs/2512.10300.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. 2025a. See what you are told: Visual attention sink in large multimodal models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. 2025b. Your large vision-language model only needs A few attention heads for visual grounding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 9339–9350. Computer Vision Foundation / IEEE.

Junho Kim, Hyunjun Kim, Yeonju Kim, and Yong Man Ro. 2024. CODE: contrasting self-generated description to combat hallucination in large multi-modal models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13872–13882. IEEE.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2020. What does BERT with vision look at? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5265–5275. Association for Computational Linguistics.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 292–305. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, Lecture Notes in Computer Science, pages 740–755. Springer.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024a. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024b. A survey on hallucination in large vision-language models. *CoRR*, abs/2402.00253.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024c. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26286–26296. IEEE.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Shi Liu, Kecheng Zheng, and Wei Chen. 2024d. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXIII*, Lecture Notes in Computer Science, pages 125–140. Springer.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4035–4045. Association for Computational Linguistics.

Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. 2016. Grounding of textual phrases in images by reconstruction. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, Lecture Notes in Computer Science, pages 817–834. Springer.

Sreetama Sarkar, Yue Che, Alex Gavin, Peter Anthony Beerel, and Souvik Kundu. 2025. Mitigating hallucinations in vision-language models through image-guided head suppression. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 12481–12500. Association for Computational Linguistics.

Mong Yuan Sim, Wei Emma Zhang, Xiang Dai, and Biaoyan Fang. 2025. Can vlms actually see and read? A survey on modality collapse in vision-language models. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Findings of ACL, pages 24452–24470. Association for Computational Linguistics.

Qwen Team. 2025. Qwen3-vl technical report. *CoRR*, abs/2511.21631.

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Iyer, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, Xichen Pan, Rob Fergus, Yann LeCun, and Saining Xie. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Robin Viltoriano, Wei Emma Zhang, Hu Wang, Mong Yuan Sim, and Yanjun Shu. 2026. Peacap: Patch-level retrieval for lightweight retrieval-augmented image captioning. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2026, MelbourneVICAustralia, July 20-24, 2026*, pages 4216–4220. ACM.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 11–20. Association for Computational Linguistics.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, and 8 others. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *CoRR*, abs/2412.10302.

Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. 2025. Understanding and mitigating hallucination in large vision-language models via modular attribution and intervention. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Hao Yin, Guangzong Si, and Zilei Wang. 2025. Clearsight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 14625–14634. Computer Vision Foundation / IEEE.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. Modeling context in referring expressions. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*, Lecture Notes in Computer Science, pages 69–85. Springer.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. 2025. IBD: alleviating hallucinations in large vision-language models via image-biased decoding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2025, Nashville, TN, USA, June 11-15, 2025*, pages 1624–1633. Computer Vision Foundation / IEEE.

A MME Evaluation Metric

MME (Fu et al., 2023) evaluates LVLMs using binary yes/no questions. For each image, MME provides two manually designed questions, where one answer is "yes" and the other is "no". Following the official MME protocol, we report both Accuracy and Accuracy$^+$. Accuracy is computed at the question level, measuring the proportion of individual yes/no questions answered correctly. In contrast, Accuracy$^+$ is computed at the image level and counts an image as correct only if the model answers both paired questions correctly. Therefore, Accuracy$^+$ is a stricter metric that better reflects whether the model consistently understands the visual content rather than guessing or showing a yes/no bias. For each subtask, the MME score is calculated as the sum of Accuracy and Accuracy$^+$, with a maximum score of 200. The overall perception and cognition scores are obtained by summing the scores over their corresponding subtasks.

B MME Evaluation Results

Table 8 presents the complete category-wise results on the MME benchmark, which evaluates both perception and cognition abilities of LVLMs. Across the three evaluated LVLMs, our method preserves the general multimodal capability of the baseline models while providing slight overall improvements. For LLaVA-1.5-7B, our method improves the perception score from 1476.8 to 1486.0, with a comparable cognition score. For Qwen3-VL, the perception score remains largely unchanged, while the cognition score improves from 665.0 to 670.0. For Deepseek2-VL, our method improves the perception score from 1665.2 to 1672.5, with only a minor decrease in cognition. We also observe improvements in OCR for both LLaVA-1.5-7B and Qwen3-VL, suggesting that the identified

Model	Method	Perception											Cognition				
		Exist.	Count	Position	Color	Posters	Celeb.	Scene	Land.	Art.	OCR	Total	Common.	Num.	Text	Code	Total
LLaVA-1.5 (7B)	Baseline	195.0	158.3	**113.3**	160.0	152.7	139.4	162.3	162.3	123.5	110.0	1476.8	120.0	**52.5**	80.0	57.5	**310.0**
	Ours	195.0	158.3	111.7	160.0	152.7	**141.8**	**163.0**	162.3	**123.8**	**117.5**	**1486.0**	120.0	52.0	80.0	57.5	309.5
Qwen3-VL	Baseline	195.0	**173.3**	143.3	190.0	181.6	**176.2**	**155.0**	**177.0**	159.0	162.5	**1713.0**	155.0	170.0	155.0	185.0	665.0
	Ours	195.0	170.0	**146.7**	190.0	181.6	175.3	151.0	169.8	**161.5**	**170.0**	1710.8	**160.0**	170.0	155.0	185.0	**670.0**
Deepseek2-VL	Baseline	195.0	138.3	**156.7**	175.0	**175.5**	177.7	161.0	175.0	141.0	170.0	1665.2	**141.4**	102.5	162.5	95.0	**501.4**
	Ours	195.0	**141.7**	151.7	175.0	174.8	**178.5**	**161.8**	**176.0**	**148.0**	170.0	**1672.5**	135.7	**105.0**	162.5	95.0	498.2

Table 8: Results on MME evaluation. Exist., Land., Art., Common., and Num. denote existence, landmark, artwork, commonsense reasoning, and numerical calculation, respectively.

visual-sensitive heads may capture localized visual evidence, including image regions containing text. Overall, these results indicate that the proposed head-level intervention does not compromise general multimodal ability, while offering modest gains on a comprehensive multimodal benchmark.

C Extended Ablation Studies

We ablate three hyperparameters of visual-gated hallucination-head suppression: the number of hallucination heads selected for intervention (top-k), the suppression factor, and the visual activation threshold. The visual threshold controls when suppression is applied: suppression is triggered only when the selected visual heads assign sufficient attention mass to image tokens. Thus, the threshold acts as a signal indicator for whether the model is currently relying on visual evidence, rather than applying hallucination-head suppression unconditionally.

Across all three models, we observe a clear precision-recall tradeoff. More aggressive settings, such as lower visual thresholds, lower suppression factors, or larger top-k, often reduce CHAIR scores, but can also reduce recall, F1, or produce abnormal caption lengths. This indicates that the intervention can remove hallucinated object mentions, but if applied too aggressively, it may also suppress visually grounded content. Therefore, we evaluate each setting not only by CHAIR_S and CHAIR_I, but also by recall, precision, F1, and caption length.

The visual threshold is useful as a gating signal because it prevents suppression from becoming always active. Low thresholds tend to fire too frequently, which can yield low CHAIR scores at the cost of reduced recall and F1. Very high thresholds fire too rarely and recover hallucination. Intermediate thresholds usually provide the best balance, suggesting that visual attention mass is informative for deciding when hallucination-head suppression should be applied.

The optimal top-k differs substantially across models. LLaVA benefits from moderate intervention sizes, with top-k 50 giving a strong tradeoff between hallucination reduction and preserved F1. Qwen3-VL is more sensitive to large intervention sets: top-k 100 or 200 can sharply reduce CHAIR, but often collapses recall and F1, suggesting oversuppression. DeepSeek-VL2-Small, in contrast, benefits most from top-k 200, where the method substantially reduces CHAIR while preserving competitive F1 under intermediate thresholds. This shows that the distribution of hallucination-related heads is architecture-dependent, and top-k should be tuned per model rather than scaled directly with model size.

The suppression factor further modulates this tradeoff. Strong suppression can improve precision and lower CHAIR, but it becomes harmful when combined with large top-k or low visual thresholds. Weak suppression preserves recall but often behaves too close to the unmodified model. In practice, moderate suppression factors around 0.3-0.5 and intermediate visual thresholds around 0.3-0.7 provide the most reliable operating region, with the exact top-k selected per model.

D Baseline Hyperparameters

Table 12 summarizes the hyperparameters used for all baseline methods in our experiments. To ensure a fair comparison, we keep the dataset split, sample range, decoding setup, and common inference settings consistent across methods whenever possible. For method-specific hyperparameters, we adopt the official recommended configurations from the corresponding papers and the released official implementations. Otherwise, we use the default values from the public codebase, or select the closest comparable settings for the corresponding model family when model-specific adjustment is necessary.

E Case Studies

We present qualitative examples to illustrate both the benefits and trade-offs of our intervention method. In our success cases, the baseline model generates object mentions that are not supported by the visual content, whereas our method suppresses these hallucinated objects while retaining the principal grounded description of the image. These examples demonstrate that the proposed intervention method can effectively reduce the reliance on language priors and encourage generation that is better aligned with visual evidence.

We additionally report failure cases in which our intervention method makes the model becomes more conservative. In the examples shown in Table 14, the baseline includes plausible or visually supported object mentions, but our method suppresses them during generation. Although this behavior can reduce hallucination risk, it may also remove some fine-grained details. These cases highlight the trade-off between recall and precision in hallucination mitigation, where improving precision by suppressing unsupported content can come at the cost of recall when the intervention is applied too strongly.

F LLaVA-Bench (In-the-Wild) Evaluation Prompt

To evaluate the performance of LVLMs on LLaVA-Bench (In-the-Wild), we use GPT-5-mini as the evaluator. The prompt template adapted (He et al., 2025; Liu et al., 2024c) is shown in Table 15. Specifically, we focus on visual accuracy, relevance and details generated from the LVLMs.

G Motivating Analysis: Dynamic Selection, Static Heads

Prior studies that detect visually sensitive heads often select heads dynamically according to their attention mass over visual tokens (Sarkar et al., 2025; Kang et al., 2025b), or head sensitivity with and without visual tokens (He et al., 2025). Our analysis finds that the selected heads are not fully sample-specific. As visualized in Figure 5, a subset of heads is repeatedly selected across samples, ranging from only a few heads in LocalizationHead (Kang et al., 2025b) to hundreds of heads in SPIN (Sarkar et al., 2025) and Vision-aware Head Divergence (VHD) (He et al., 2025). This indicates that dynamic selection may implicitly rely on stable, visually sensitive heads. Motivated by this, we explicitly detect such heads and use their attention behavior as a step-level signal for when visual grounding is needed.

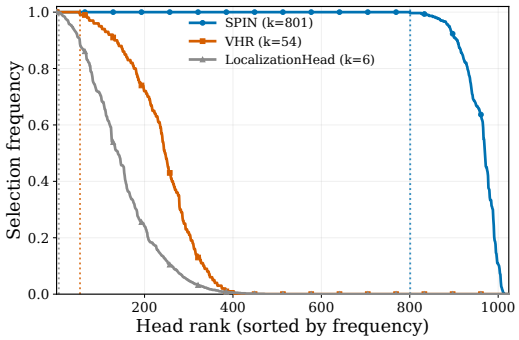


Figure 5: Stable head selection in existing dynamic head-selection methods. Although these methods select heads per sample or decoding step, there exists a subset of heads that is repeatedly selected across the dataset.

Top-k	Suppress Factor	Visual Threshold	$CHAIR_S$	$CHAIR_I$	Recall	Precision	F1	Len
25	0.0	0.0	25.8	7.8	65.5	85.5	74.2	72.0
		0.1	25.3	8.0	65.7	85.8	74.4	72.5
		0.3	26.7	8.3	65.7	84.8	74.1	72.7
		0.5	26.8	8.4	66.5	84.8	74.5	73.8
		0.7	28.9	9.2	67.0	83.5	74.3	74.4
		0.9	31.0	9.8	67.8	82.7	74.5	74.8
	0.3	0.0	29.0	8.9	66.6	83.7	74.2	73.4
		0.1	29.6	8.9	66.6	83.8	74.2	73.0
		0.3	29.9	9.3	66.8	83.2	74.1	74.0
		0.5	30.4	9.6	67.3	82.7	74.2	74.5
		0.7	30.7	9.9	67.7	83.1	74.6	75.1
		0.9	31.9	9.9	68.3	82.8	74.9	75.0
	0.5	0.0	30.1	9.1	67.7	83.6	74.8	73.9
		0.1	30.4	9.5	67.9	83.4	74.9	74.2
		0.3	30.7	9.8	67.8	83.2	74.7	74.5
		0.5	31.0	10.0	68.2	83.2	74.9	74.5
		0.7	32.0	10.5	68.4	82.5	74.8	74.8
		0.9	32.9	10.2	68.3	82.2	74.6	74.6
	0.9	0.0	32.1	10.2	68.5	82.2	74.7	75.1
		0.1	32.3	10.2	68.6	82.3	74.8	75.1
		0.3	32.4	10.2	68.5	82.2	74.7	75.0
		0.5	32.3	10.3	68.4	82.1	74.6	75.0
		0.7	31.8	10.4	68.3	81.9	74.5	74.8
		0.9	31.5	10.4	68.3	81.8	74.5	75.1
50	0.0	0.0	23.2	6.2	64.1	87.7	74.1	80.8
		0.1	23.6	6.9	64.1	87.6	74.1	78.1
		0.3	25.0	7.8	65.0	86.4	74.6	76.3
		0.5	27.4	7.7	66.2	85.4	74.2	76.8
		0.7	29.2	8.5	66.5	84.5	74.4	75.1
		0.9	30.8	9.5	67.6	82.8	74.4	76.3
	0.3	0.0	25.2	7.9	65.7	86.2	74.6	76.0
		0.1	25.3	8.3	66.3	86.2	74.9	76.2
		0.3	26.7	7.8	67.1	85.6	75.3	75.7
		0.5	29.0	8.6	67.4	84.2	74.9	75.4
		0.7	30.6	9.5	67.3	82.9	74.3	76.1
		0.9	31.5	9.7	68.1	82.5	74.6	75.7
	0.5	0.0	27.5	8.2	67.0	85.3	75.1	75.7
		0.1	28.3	8.3	67.4	85.1	75.2	75.6
		0.3	28.7	8.9	67.6	84.3	75.0	75.5
		0.5	29.9	9.1	67.8	83.8	74.9	76.2
		0.7	32.0	9.9	67.7	82.5	74.4	75.9
		0.9	31.0	9.7	68.2	82.9	74.9	75.3
	0.9	0.0	32.0	9.9	68.6	82.6	74.9	75.8
		0.1	31.8	9.9	68.4	82.4	74.8	75.7
		0.3	31.9	9.9	68.6	82.7	75.0	75.6
		0.5	32.6	10.1	68.7	82.2	74.8	75.5
		0.7	31.8	10.1	68.7	82.5	75.0	75.2
		0.9	31.8	10.2	68.4	82.2	74.6	75.0
100	0.0	0.0	0.6	0.5	9.9	98.0	17.9	29.8
		0.1	1.1	1.9	6.8	95.1	12.8	84.0
		0.3	9.5	6.1	41.1	92.3	56.8	177.0
		0.5	14.1	6.3	45.1	88.4	59.8	156.1
		0.7	25.1	7.6	62.8	85.2	72.3	74.9
		0.9	29.4	9.1	67.7	83.2	74.6	76.0
	0.3	0.0	19.9	7.5	57.3	88.3	69.5	104.1
		0.1	20.1	7.9	57.9	88.5	70.0	102.5
		0.3	22.8	7.0	60.3	87.0	71.2	89.9
		0.5	25.9	9.6	61.9	84.6	71.5	85.8
		0.7	27.8	8.8	64.8	84.1	73.2	73.1
		0.9	31.2	9.5	68.0	82.5	74.6	75.6
	0.5	0.0	25.5	7.6	63.6	86.4	73.3	77.3
		0.1	25.6	7.6	63.9	86.6	73.5	75.2
		0.3	29.1	8.5	64.6	84.7	73.3	73.0
		0.5	28.7	8.7	65.9	84.2	74.0	73.2
		0.7	29.3	9.2	66.4	83.3	73.9	73.1
		0.9	30.4	9.4	68.6	83.0	75.1	75.2
	0.9	0.0	30.6	9.3	67.5	83.4	74.6	75.5
		0.1	30.7	9.3	68.0	83.5	75.0	75.6
		0.3	32.0	9.8	67.5	82.5	74.3	74.9
		0.5	31.9	9.8	68.0	82.8	74.7	74.6
		0.7	32.2	9.9	68.5	82.5	74.9	75.0
		0.9	32.2	10.2	68.7	82.1	74.8	75.1
200	0.0	0.0	5.1	12.5	40.3	96.0	56.8	86.6
		0.1	4.0	7.2	41.4	97.1	58.1	69.9
		0.3	6.4	4.2	45.2	95.6	61.4	55.8
		0.5	11.7	5.1	51.5	92.3	66.1	94.1
		0.7	22.7	7.2	61.3	86.7	71.8	71.7
		0.9	28.0	8.5	66.2	84.7	74.4	74.4
	0.3	0.0	17.8	7.4	57.5	90.0	70.1	122.8
		0.1	18.7	6.0	58.4	89.4	70.7	112.0
		0.3	20.2	6.6	62.5	89.0	73.5	92.5
		0.5	21.8	5.7	64.2	88.5	74.4	78.7
		0.7	27.5	8.1	67.5	85.1	75.2	75.9
		0.9	30.2	9.5	67.8	83.1	74.6	75.2
	0.5	0.0	24.7	7.7	64.0	86.7	73.7	84.9
		0.1	24.7	8.0	64.7	86.8	74.1	82.6
		0.3	27.2	7.9	65.7	85.7	74.3	77.0
		0.5	27.6	8.0	66.4	85.4	74.7	75.3
		0.7	30.5	8.7	67.4	83.5	74.6	76.1
		0.9	30.9	9.8	68.5	82.7	74.9	75.4
	0.9	0.0	31.9	9.4	68.2	83.2	74.9	76.1
		0.1	32.0	9.6	68.2	82.7	74.8	75.9
		0.3	32.2	9.8	68.2	82.6	74.7	75.5
		0.5	31.6	9.9	68.5	82.8	75.0	75.0
		0.7	31.9	10.2	68.5	82.3	74.8	74.8
		0.9	32.0	10.3	68.7	81.8	74.7	75.4

Table 9: CHAIR ablation results for LLaVA-1.5-7B. Metrics are percentages except Len, which is the average caption length in tokens. Rows highlighted in red indicate severe degenerate behavior, such as collapsed F1 or abnormal caption length.

Top-k	Suppress Factor	Visual Threshold	$CHAIR_s$	$CHAIR_i$	Recall	Precision	F1	Len
25	0.0	0.0	52.7	9.5	73.5	74.9	74.2	387.2
		0.1	51.7	9.7	73.6	75.0	74.3	381.5
		0.3	51.9	9.7	74.8	74.9	74.8	381.2
		0.5	52.9	10.2	75.0	74.7	74.8	380.2
		0.7	55.1	10.4	75.5	73.9	74.7	382.0
		0.9	54.0	10.4	76.4	74.1	75.2	384.6
	0.3	0.0	55.5	9.8	75.3	73.9	74.6	385.9
		0.1	53.4	9.5	75.7	74.6	75.1	384.3
		0.3	53.1	9.9	75.0	74.8	74.9	384.0
		0.5	53.7	10.1	75.9	74.3	75.1	384.7
		0.7	53.4	10.3	75.7	74.6	75.2	382.1
		0.9	54.6	10.4	76.6	73.6	75.0	384.4
	0.5	0.0	54.5	10.1	76.0	74.3	75.1	386.6
		0.1	54.1	10.1	75.7	74.3	75.0	385.3
		0.3	55.4	10.2	75.9	73.9	74.9	386.6
		0.5	55.3	10.3	76.0	74.0	75.0	385.5
		0.7	55.3	10.5	75.8	73.6	74.7	383.3
		0.9	54.1	10.7	76.1	73.8	74.9	384.6
	0.9	0.0	55.3	10.3	76.2	73.7	74.9	384.6
		0.1	54.3	10.5	75.9	73.7	74.8	385.1
		0.3	55.2	10.7	76.5	73.0	74.7	384.9
		0.5	53.2	10.4	76.2	74.2	75.2	385.2
		0.7	56.0	10.6	76.2	73.6	74.9	384.8
		0.9	57.7	10.7	76.9	72.7	74.7	384.2
50	0.0	0.0	38.2	6.7	65.7	80.7	72.4	440.2
		0.1	38.1	6.3	66.1	80.7	72.7	439.0
		0.3	45.6	7.4	69.5	78.1	73.6	434.8
		0.5	49.3	8.3	72.8	76.2	74.5	427.9
		0.7	52.8	9.3	75.3	74.6	74.9	411.3
		0.9	54.1	10.4	76.6	73.7	75.1	390.7
	0.3	0.0	51.1	8.3	74.1	75.6	74.9	437.8
		0.1	49.4	8.1	73.5	76.2	74.8	437.0
		0.3	51.4	9.0	74.2	74.8	74.5	432.3
		0.5	54.3	9.6	74.5	74.2	74.4	420.5
		0.7	53.2	9.7	75.8	73.9	74.9	404.0
		0.9	52.6	10.5	75.7	73.9	74.8	388.3
	0.5	0.0	53.3	9.2	75.4	74.2	74.8	424.7
		0.1	53.1	8.9	75.5	74.6	75.1	427.1
		0.3	52.5	9.4	75.3	74.2	74.8	418.7
		0.5	53.3	9.7	75.8	74.4	75.1	410.3
		0.7	55.8	10.3	75.9	73.1	74.5	401.5
		0.9	53.8	10.8	75.5	73.4	74.5	386.0
	0.9	0.0	53.6	10.2	76.3	74.2	75.2	394.4
		0.1	55.7	10.3	76.1	73.5	74.8	393.2
		0.3	55.4	10.5	76.2	73.4	74.8	391.5
		0.5	55.0	10.6	75.9	73.5	74.7	391.5
		0.7	55.7	10.9	76.2	73.2	74.7	388.3
		0.9	56.1	10.3	76.1	73.4	74.7	382.8
100	0.0	0.0	0.7	6.1	4.7	95.5	8.9	578.3
		0.1	1.2	2.6	9.5	96.2	17.3	533.1
		0.3	4.2	4.1	19.8	93.3	32.6	503.1
		0.5	8.1	6.0	29.0	91.6	44.1	487.3
		0.7	14.6	6.5	40.1	88.2	55.1	466.8
		0.9	26.9	7.1	55.1	83.9	66.5	430.6
	0.3	0.0	3.4	3.4	19.5	94.7	32.3	513.8
		0.1	5.9	4.9	25.0	92.6	39.3	504.6
		0.3	9.2	5.4	32.3	90.7	47.6	492.0
		0.5	15.2	5.9	40.4	88.3	55.4	475.0
		0.7	23.0	6.5	48.4	84.7	61.6	444.0
		0.9	31.2	7.5	58.4	81.0	67.9	418.2
	0.5	0.0	10.4	4.0	33.0	90.1	48.3	496.9
		0.1	11.4	4.4	36.0	90.0	51.5	493.8
		0.3	17.6	5.5	42.6	86.7	57.1	475.0
		0.5	23.1	6.3	49.4	84.9	62.5	453.9
		0.7	27.8	6.4	54.8	83.5	66.2	438.4
		0.9	35.2	9.1	60.1	80.4	68.8	414.1
	0.9	0.0	34.1	9.2	59.7	79.9	68.3	433.9
		0.1	35.1	9.6	60.0	79.8	68.5	429.1
		0.3	34.6	8.7	60.1	80.1	68.7	426.1
		0.5	35.4	9.1	61.2	80.0	69.3	419.3
		0.7	36.7	9.3	62.1	79.1	69.6	409.3
		0.9	37.0	9.0	63.1	79.2	70.3	412.0
200	0.0	0.0	0.1	20.7	0.2	88.9	0.5	466.1
		0.1	0.0	0.0	0.4	100.0	0.9	457.2
		0.3	1.1	6.8	5.6	93.8	10.6	494.5
		0.5	3.0	3.0	16.6	94.5	28.2	495.9
		0.7	9.2	4.8	30.8	90.6	45.9	489.3
		0.9	26.4	5.6	52.6	83.9	64.7	446.0
	0.3	0.0	0.6	7.2	8.0	97.7	14.8	449.4
		0.1	1.0	3.8	11.3	97.1	20.2	466.9
		0.3	4.2	5.5	21.0	93.8	34.3	493.3
		0.5	9.6	4.8	31.1	90.3	46.2	488.9
		0.7	17.7	6.6	40.9	86.4	55.6	469.4
		0.9	29.9	9.0	57.9	82.3	68.0	431.2
	0.5	0.0	5.5	2.1	27.1	93.6	42.0	483.7
		0.1	7.1	2.9	29.6	92.5	44.8	492.5
		0.3	10.8	5.1	35.5	90.5	51.0	483.1
		0.5	18.4	5.0	43.4	87.0	57.9	465.2
		0.7	23.9	6.8	51.0	84.2	63.5	447.8
		0.9	33.2	7.5	59.3	80.3	68.2	421.9
	0.9	0.0	33.7	8.0	60.3	80.9	69.1	429.6
		0.1	33.0	6.8	59.7	80.8	68.7	426.2
		0.3	35.9	7.7	61.2	79.6	69.2	418.0
		0.5	36.0	8.6	61.1	78.8	68.8	414.1
		0.7	36.6	10.7	61.9	79.4	69.5	411.6
		0.9	35.0	7.6	62.9	80.2	70.5	407.6

Table 10: CHAIR ablation results for Qwen3-VL-8B. Metrics are percentages except Len, which is the average caption length in tokens. Rows highlighted in red indicate severe degenerate behavior, such as collapsed F1 or abnormal caption length.

Top-k	Suppress Factor	Visual Threshold	$CHAIR_s$	$CHAIR_i$	Recall	Precision	F1	Len
25	0.0	0.0	47.1	13.3	68.3	76.9	72.3	183.2
		0.1	47.1	13.4	68.8	76.7	72.6	184.3
		0.3	48.0	14.5	69.6	75.5	72.4	186.5
		0.5	48.5	14.3	69.4	76.2	72.6	186.1
		0.7	50.2	14.5	69.8	75.9	72.7	185.1
		0.9	50.7	14.9	70.1	75.4	72.7	185.8
	0.3	0.0	47.3	13.5	69.5	76.8	73.0	182.7
		0.1	47.9	13.7	68.4	76.4	72.2	183.9
		0.3	46.3	14.5	69.0	76.1	72.4	183.9
		0.5	49.6	14.4	69.5	75.9	72.6	186.4
		0.7	51.1	15.0	69.3	75.0	72.0	185.4
		0.9	50.9	14.7	69.1	75.5	72.2	185.9
	0.5	0.0	48.2	13.6	70.2	76.6	73.2	183.5
		0.1	49.7	14.1	69.4	75.9	72.5	183.6
		0.3	47.3	14.5	69.8	76.4	72.9	184.5
		0.5	49.2	14.6	69.1	76.0	72.4	184.8
		0.7	50.9	14.7	69.9	75.8	72.7	184.3
		0.9	48.7	14.6	69.6	76.0	72.7	185.2
	0.9	0.0	49.3	14.0	69.3	76.5	72.7	185.0
		0.1	48.7	14.4	69.6	75.9	72.6	184.9
		0.3	50.2	14.6	69.9	75.9	72.7	185.4
		0.5	50.3	14.5	70.3	76.0	73.0	185.3
		0.7	49.7	14.8	69.3	75.7	72.3	184.9
		0.9	50.4	14.7	69.3	75.9	72.5	185.6
50	0.0	0.0	37.0	12.2	64.7	82.0	72.3	216.1
		0.1	40.0	10.7	66.3	80.8	72.8	195.1
		0.3	42.9	11.8	67.7	79.1	73.0	178.0
		0.5	48.7	13.5	69.5	76.9	73.0	181.5
		0.7	49.8	14.3	69.3	75.8	72.4	182.6
		0.9	52.2	15.5	69.8	74.4	72.0	184.8
	0.3	0.0	42.9	12.8	66.7	78.5	72.1	182.8
		0.1	46.7	12.2	68.2	78.0	72.8	182.9
		0.3	49.6	13.7	68.8	76.2	72.3	183.4
		0.5	48.4	14.1	69.5	76.0	72.6	182.7
		0.7	49.1	14.6	69.3	75.7	72.4	182.1
		0.9	51.5	15.2	69.6	75.0	72.2	184.8
	0.5	0.0	45.5	12.5	68.4	78.1	73.0	179.7
		0.1	46.8	12.8	68.7	77.8	73.0	181.2
		0.3	49.3	13.8	70.0	76.7	73.2	181.4
		0.5	48.7	13.9	68.6	76.2	72.2	183.4
		0.7	50.4	14.7	69.4	75.3	72.2	183.7
		0.9	50.1	14.9	69.5	75.4	72.3	184.9
	0.9	0.0	50.6	14.5	69.8	75.9	72.7	185.0
		0.1	51.4	14.9	69.1	75.3	72.1	184.4
		0.3	51.4	14.7	69.5	75.4	72.3	184.6
		0.5	49.2	14.4	69.2	75.9	72.4	184.9
		0.7	49.0	14.5	69.7	75.9	72.6	184.7
		0.9	50.3	14.7	69.3	75.6	72.3	185.7
100	0.0	0.0	21.8	7.3	54.6	87.7	67.3	396.0
		0.1	26.1	10.5	57.2	85.9	68.7	284.6
		0.3	43.8	11.8	67.6	78.8	72.8	188.5
		0.5	47.7	13.5	69.5	77.3	73.2	179.4
		0.7	48.3	14.0	68.5	76.5	72.3	181.1
		0.9	50.5	14.7	69.7	75.6	72.5	183.3
	0.3	0.0	40.2	12.1	66.1	80.0	72.4	185.2
		0.1	38.5	10.7	66.3	81.0	72.9	182.6
		0.3	45.7	12.7	68.4	78.1	73.0	181.0
		0.5	48.0	14.2	69.0	76.2	72.4	180.5
		0.7	47.6	14.0	69.6	76.5	72.9	182.7
		0.9	49.3	14.7	69.9	75.6	72.6	183.8
	0.5	0.0	42.9	11.8	67.7	79.1	73.0	178.0
		0.1	45.0	12.2	67.7	78.3	72.6	177.6
		0.3	46.1	13.4	69.1	77.6	73.1	179.2
		0.5	47.9	14.1	69.7	76.5	72.9	182.9
		0.7	50.4	14.6	69.4	75.4	72.3	182.5
		0.9	51.4	15.1	69.1	75.3	72.1	184.7
	0.9	0.0	49.1	14.2	69.4	76.2	72.6	183.4
		0.1	49.4	14.2	69.2	76.2	72.5	182.9
		0.3	49.1	14.2	69.0	75.9	72.3	185.2
		0.5	49.1	14.5	68.9	75.8	72.2	183.1
		0.7	48.5	14.5	69.2	75.8	72.3	186.2
		0.9	49.8	14.6	69.8	75.8	72.7	186.1
200	0.0	0.0	5.8	1.1	50.1	96.4	65.9	395.5
		0.1	6.9	1.8	52.5	95.9	67.9	366.6
		0.3	17.4	4.1	61.6	91.1	73.5	169.8
		0.5	22.7	5.7	65.5	88.7	75.4	115.5
		0.7	26.4	6.6	67.7	87.1	76.2	127.0
		0.9	48.1	14.2	69.1	76.2	72.5	181.2
	0.3	0.0	16.6	5.2	60.8	91.3	73.0	122.7
		0.1	17.3	4.9	62.4	91.2	74.1	117.6
		0.3	20.3	6.0	64.8	89.5	75.2	110.4
		0.5	24.0	6.1	67.2	88.4	76.3	110.1
		0.7	28.6	7.0	67.9	86.2	76.0	137.3
		0.9	50.2	14.4	69.5	75.9	72.5	182.2
	0.5	0.0	19.8	5.1	64.8	90.3	75.4	96.0
		0.1	21.1	5.6	64.8	89.2	75.0	102.1
		0.3	24.7	6.1	67.2	87.9	76.2	115.5
		0.5	26.9	6.5	68.1	87.0	76.4	122.2
		0.7	36.3	9.9	69.3	81.6	74.9	159.5
		0.9	48.5	14.5	69.3	75.8	72.4	183.9
	0.9	0.0	49.9	14.0	68.3	75.8	71.9	178.1
		0.1	49.2	14.0	68.8	76.1	72.3	179.1
		0.3	49.5	14.1	68.9	75.8	72.2	181.5
		0.5	49.2	14.0	69.5	76.6	72.8	182.8
		0.7	50.1	14.5	69.6	75.8	72.6	185.2
		0.9	48.7	14.5	69.4	76.0	72.6	184.9

Table 11: CHAIR ablation results for DeepSeek-VL2-Small. Metrics are percentages except Len, which is the average caption length in tokens. Rows highlighted in red indicate severe degenerate behavior, such as collapsed F1 or abnormal caption length.

Model	Method	Hyperparameter	Value
LLaVA-1.5-7B	PAI	alpha	0.2
		use_cfg	off
	DAMRO	alpha_damro	0.2
		outlier_topk	10
	ADHH	adhh_topk	20
		adhh_threshold	0.0
	VCD	noise_step	999
		cd_alpha	1.0
		cd_beta	0.1
	OPERA	beam	1
		scale_factor	50.0
		threshold	15
		num_candidates	5
		penalty_weights	1.0
	SPIN	routed_heads	0.95
	VHR	vhr_aug_ratio	2.0
Qwen3-VL-8B	PAI	start_layer	35
		end_layer	36
		alpha	0.005
		use_cfg	off
	DAMRO	alpha_damro	0.2
		outlier_topk	10
	ADHH	start_layer	0
		end_layer	36
		adhh_topk	20
		adhh_threshold	0.0
	VCD	noise_step	999
		cd_alpha	1.0
		cd_beta	0.1
	OPERA	beam	5
		scale_factor	5.0
		threshold	100
		num_candidates	5
		penalty_weights	0.1
	SPIN	routed_heads	0.95
		small_num_mask	0.5
		manual_topk	unset
	VHR	vhr_aug_ratio	0.5
		vhr_last_layers	4
		vhr_filter	on
DeepSeek-VL2-Small	PAI	alpha	0.2
		use_cfg	off
	DAMRO	alpha_damro	0.2
		outlier_topk	10
	ADHH	adhh_topk	20
		adhh_threshold	0.0
	VCD	noise_step	999
		cd_alpha	1.0
		cd_beta	0.1
	OPERA	scale_factor	50.0
		threshold	15
		num_candidates	5
		penalty_weights	1.0
	SPIN	routed_heads	0.95
		small_num_mask	unset
		manual_topk	unset
	VHR	vhr_aug_ratio	2.0
		vhr_filter	off

Table 12: Hyperparameter settings for the seven intervention methods evaluated across three LVLM backbones. For LLaVA-1.5-7B, the VCD/OPERA/SPIN/VHR settings are taken from the benchmark implementation, while PAI, DAMRO, and ADHH are taken from the dedicated CHAIR launch scripts.

Sample 1: COCO image 253426



Qwen3-VL Greedy: "This is a close-up, slightly angled photograph of a single slice of pizza resting on an ornate plate. ... The background is out of focus, with a hint of a brown surface and what might be a **book** or menu with text visible on the right side, suggesting a casual dining setting."

Ours: "This is a close-up, top-down photograph of a single slice of pizza resting on a decorative plate. ... The background is slightly out of focus, with a glimpse of a brown surface, possibly a tablecloth or placemat, visible behind the pizza."

Sample 2: COCO image 020070



Qwen3-VL Greedy: "This is a candid, slightly humorous photograph of a man in what appears to be a bathroom or a similar personal space. ... A small blue object, possibly a **toothbrush** or a **bottle** cap, can be seen on the counter."

Ours: "This is a candid, indoor photograph of a man in what appears to be a bathroom or a similar personal space, captured in a moment of lightheartedness. ... He is holding a white hairdryer in his right hand."

Sample 3: COCO image 455633



Qwen3-VL Greedy: "This is a dynamic, candid photograph capturing a moment at a surf park or river surfing location. ... Further back, a group of people, including children, are gathered, some sitting on a **bench** or in **chairs**, watching the activity."

Our: "This is a dynamic, action-oriented photograph capturing a moment at a river or artificial wave pool. ... Behind the main subject, there is a group of people, some of whom are also holding surfboards."

Table 13: **Qualitative Success Cases of Visual Gate on Qwen3-VL.** The greedy baseline introduces object mentions that are not supported by the image (shown in **red bold**), while our visual gate suppresses them. Relevant verbatim output excerpts are shown for readability.

Sample 1: COCO image 438807



Qwen3-VL Greedy: "This is a dynamic, action-oriented photograph capturing a surfer riding a wave in the ocean under an overcast sky. ... The **surfboard**, partially visible beneath them, appears to be a light color, possibly yellow or white, and is cutting through the water."

Ours (Visual Gate): "This is a dynamic, action-oriented photograph capturing a surfer in the midst of riding a wave. ... Their body is angled low and close to the water, suggesting they are executing a maneuver or maintaining balance."

Sample 2: COCO image 571182



Qwen3-VL Greedy: "This is a moody, atmospheric photograph capturing a snowy urban scene. ... **Traffic lights** and street signs are visible, including a green arrow sign."

Ours (Visual Gate): "The image captures a dynamic urban scene during a snowy day. ... A large, bright green public bus is the most prominent moving element in the scene."

Sample 3: COCO image 217936



Qwen3-VL Greedy: "This is a dynamic, action-oriented photograph of a women's tennis match. ... The central focus is a female tennis player in the foreground, captured mid-swing as she prepares to hit the **ball**."

Ours (Visual Gate): "This is a dynamic, action-oriented photograph of a women's tennis match. ... The central focus is a female tennis player in the foreground, captured mid-stride as she lunges forward."

Table 14: **Qualitative Failure Cases of Visual Gate on Qwen3-VL: Over-Suppression.** The greedy baseline contains visually supported or contextually warranted object mentions (shown in **blue bold**) that our visual gate removes. Relevant verbatim output excerpts are shown for readability.

LLaVA-Bench (In-the-Wild) Evaluation Prompt on GPT-5-mini

You are an impartial evaluator. You will be given:

1. a user question about an image,
2. a textual description of the image as the visual reference,
3. two assistant responses.

Evaluate each assistant response according to the visual reference.
Please score each response on the following criteria from 1 to 10:

1. **visual_accuracy**: Whether the response is consistent with the visual reference and does not introduce unsupported objects, attributes, actions, counts, spatial relations, or other details.
2. **relevance**: Whether the response directly answers the user question.
3. **detail**: Whether the response provides an appropriate level of detail. More detail is not always better; unsupported or unnecessary details should be penalized.

Use integer scores from 1 to 10, where 10 is best.
Important instructions:

- Visual accuracy should be judged only according to the provided visual reference.
- Do not reward a response for adding details that are not supported by the visual reference.
- Avoid bias caused by the order in which the two responses are presented.
- Evaluate each response independently before comparing them.
- Return only valid JSON.
- Do not include markdown, comments, or any text outside the JSON object.

Return the result in the following format:

```
{
  "assistant_1": {
    "visual_accuracy": 0,
    "relevance": 0,
    "detail": 0,
    "explanation": ""
  },
  "assistant_2": {
    "visual_accuracy": 0,
    "relevance": 0,
    "detail": 0,
    "explanation": ""
  }
}
```

Table 15: The prompt used for GPT-5-mini evaluation on LLaVA-Bench (In-the-Wild).