Mitigating Hallucinations in VLMs: Enhancing Visual Attention via Head-Wise Perturbation

Zhenghua Wang [*] , **Yixin Wu** [*] , **Feiran Zhang, Qi Qian, Changze Lv, Xuanjing Huang, Xiaoqing Zheng** [†]
College of Computer Science and Artificial Intelligence, Fudan University, Shanghai, China
Shanghai Key Laboratory of Intelligent Information Processing
{zhenghuawang23, yixinwu23}@m.fudan.edu.cn
{xjhuang,zhengxq}@fudan.edu.cn

Abstract

Vision–Language Models (VLMs) have demonstrated strong capabilities in tasks that require joint understanding of text and images. However, as many VLMs are built upon pre-trained large language models, they often over-rely on linguistic priors at the expense of visual features, causing persistent hallucinations. We observe that these hallucinations stem not only from insufficient visual attention but also from imbalanced activation profiles across attention heads, while hallucinated samples tend to disproportionately activate heads that fail to capture visual cues. To promote a more balanced attention distribution, we propose **HWP**, a strategy that incorporates head-wise attention perturbation via continuous multiplicative noise, coupled with a visual-guided loss focused on vision-sensitive text tokens. Beyond simply strengthening visual grounding, this design encourages a broader set of attention heads to engage with visual signals, thereby alleviating visual information loss caused by activation concentration on a few visually blind heads. Consistent gains across different architectures and scales on multiple benchmarks demonstrate the effectiveness and robustness of our approach in mitigating VLM hallucinations.

1 Introduction

Recent advancements in vision-language models (VLMs) (Wang et al., 2024; Wei et al., 2025; Team et al., 2025) have empowered LLMs to integrate visual inputs, achieving remarkable success in tasks such as visual question answering (Chen et al., 2025b), embodied interaction (Liu et al., 2024; Yang et al., 2025; Lin et al., 2025), and tool-augmented agents (Yang et al., 2024a,b). Despite such progress, current architectures often exhibit an attentional bias, under-utilizing visual evidence and instead relying disproportionately on their dominant linguistic priors (Chen et al., 2025a; Ghatkesar et al., 2025). This imbalance frequently precipitates hallucinations—textual outputs that diverge from visual ground truth—ultimately undermining the reliability and practical utility of VLMs. These shortcomings underscore the critical need for mechanisms that can more effectively anchor model reasoning in visual content.

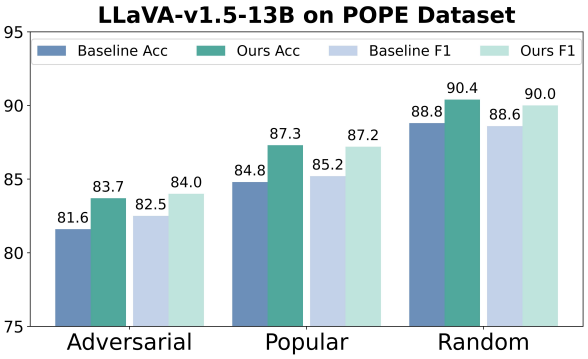

Figure 1: Performance comparison of **HWP**. On the POPE benchmark, our method consistently outperforms the baseline model, across three settings, where accuracy and F1 improve by 2.1 and 1.6 points on average.

To address the inadequate utilization of visual information in VLMs, three primary lines of research have emerged. From the **input level**, data-centric strategies (Sarkar et al., 2024; Chen et al., 2025a) aim to improve visual grounding through data augmentation. However, such methods often depend on extensive manual annotation, limiting their scalability and application. From the **model level**, several works propose redistributing attention scores to prioritize image tokens (Tang et al., 2025; Kang et al., 2025). While these methods directly target the model's internal reasoning, they rely heavily on handcrafted heuristics or static priors, raising concerns about robustness and generalizability. From the **output level**, contrastive decoding techniques have been developed to favor tokens with strong visual dependencies (Liang et al., 2024; Lee et al., 2024). While effective in emphasizing

[*]Equal contribution.
[†]Corresponding Author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 20310–20321

visual information, these methods incur significant computational overhead and require careful tuning of hyperparameters to balance contrastive strength against linguistic fluency.

Despite these efforts, VLMs still struggle to fully harness visual information, necessitating the exploration of alternative solutions. Recent studies (Wang et al., 2025; Sarkar et al., 2025) have established that only a minority subset of attention heads are functionally responsible for processing visual data. Our visualization of attention scores on various samples (Figure 3) confirms these findings, revealing that a substantial majority of heads exhibit negligible attention density toward image tokens. Meanwhile, we also identify that hallucinated samples display asymmetric activation patterns compared to faithful samples (Figure 4), characterized by weak activation in "**visual heads**" that genuinely attend to visual information, and excessive activation in "non-visual heads". This systemic imbalance leads to the erosion of visual evidence, ultimately manifesting as model hallucinations.

Building upon these insights, a straightforward remedy is to incentivize a broader set of attention heads to participate in visual processing, thereby heightening the model's sensitivity to visual stimuli, and also alleviating the degradation of visual evidence stemming from imbalanced head activation profiles. While an intuitive solution might involve applying dropout to the attention heads, standard dropout acts as multiplicative noise sampled from a Bernoulli distribution, which often introduces training instabilities due to the abrupt zeroing of activations. To circumvent this, as shown in Figure 2, we propose the integration of **multiplicative uniform noise**. This mechanism ensures stable gradient propagation while maintaining the features of dropout, promoting more heads to participate in visual processing. Furthermore, we restrict our training objective to "**visual tokens**"—specifically, text tokens exhibiting high sensitivity to visual context. These tokens are identified by calculating the logit divergence between the original image and a random-noise baseline, thereby ensuring that our optimization is aligned with visual relevance.

In summary, our contributions are three-fold:

- We introduce a novel and effective training strategy (HWP) that integrates stochastic noise injection within Multi-Head Attention (MHA), combined with a visually-grounded training scheme. The former incentivizes a broader set of attention heads to engage with visual signals, while the latter selectively prioritizes text tokens with high visual correspondence, effectively mitigating the underutilization of visual cues inherent in VLMs.
- We provide an in-depth analysis of attention dynamics, revealing that pathological sparsity in visual attention and anomalous activation profiles across attention heads are influential drivers of hallucinations. These findings offer the community new mechanistic insights into the relationship between internal attention allocation and model reliability.
- We conduct extensive experiments across diverse vision-language datasets and model architectures, demonstrating that our approach significantly enhances visual grounding and achieves a substantial reduction in hallucinatory outputs.

2 Related Work

2.1 VLM Attention Heads

The Multi-Head Attention (MHA) mechanism is inherently designed to capture diverse relational patterns within input sequences. However, recent studies (Wang et al., 2025; Bi et al., 2025; Kim et al., 2025; Deng and Yang, 2025) suggest that functional engagement with visual information in VLMs is concentrated within a small subset of these heads. For instance, Wang et al. (2025) categorize visual attention heads by evaluating their key-region localization capabilities in OCR-centric tasks, revealing that approximately less than 5% of all attention heads contribute substantially to visual understanding. Deng and Yang (2025) propose masking these visual heads to amplify the linguistic priors of VLMs, and subsequently utilize a contrastive decoding strategy to penalize these priors, thereby encouraging stronger grounding in visual content. Similarly, Sarkar et al. (2025) further highlight the importance of these specialized genuine visual heads by suppressing those dominated by linguistic priors.

2.2 Dropout and Multiplicative Noise

Dropout (Srivastava et al., 2014) is a ubiquitous regularization technique in neural networks that stochastically zeros a subset of activations during training to mitigate feature co-adaptation and encourage representation redundancy. From a the-

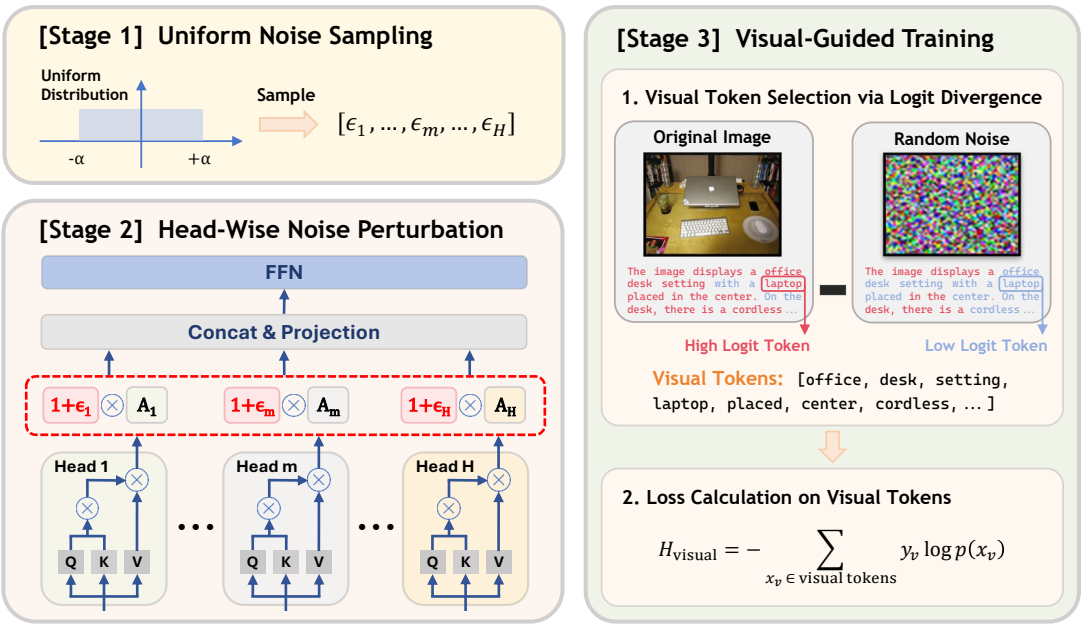

Figure 2: **Overview of HWP.** For each attention head within a Transformer layer, a noise scalar is sampled from a uniform distribution $\mathcal{U}(-a, a)$, added to 1, and multiplied with the head outputs during the forward pass. During training, we identify text tokens with high visual correlation according to their logit divergence between original and random-noise inputs. These "visual tokens" are prioritized for loss calculation, compelling focus upon visual cues.

oretical perspective, dropout can be interpreted as a form of multiplicative noise (Nalisnick et al., 2015; Li and Liu, 2016; Shen et al., 2017). These studies demonstrate that sampling multiplicative noise from continuous distributions (e.g., uniform or Gaussian), rather than a discrete Bernoulli distribution, can facilitate a smoother optimization landscape and provide more stable gradient flow with reduced variance.

3 Preliminaries

3.1 Vision-Language Models

Vision-Language Models (VLMs) augment Large Language Models (LLMs) by integrating the capability to interpret and reason over visual inputs, generally comprising a vision encoder, a cross-modal projector, and a backbone LLM. Given a visual input $\mathbf{I}$ and a sequence of text tokens with embeddings $\mathbf{E}_t$, a VLM first extracts visual features via a vision encoder and projects them into the LLM embedding space. The projected visual representation is then concatenated with the text token embeddings and fed into the backbone LLM for multimodal reasoning:

$$\mathbf{O} = \text{LLM}([\text{Projector}(\text{VisionEncoder}(\mathbf{I})); \mathbf{E}_t]), \quad (1)$$

where $[\cdot;\cdot]$ denotes concatenation, and $\mathbf{O}$ represents the resulting output from the LLM.

3.2 Multi-Head Attention

Multi-Head Attention (MHA) (Vaswani et al., 2017) serves as a foundational mechanism of modern Transformer architectures, enabling the model to concurrently attend to information from distinct representation subspaces at multiple positions. Given an input sequence of token embeddings $X \in \mathbb{R}^{T\times d}$, where T denotes the sequence length and d represents the model dimensionality, the MHA mechanism projects X into H parallel attention heads via learnable weight matrices:

$$Q_h = X \cdot W_h^Q, K_h = X \cdot W_h^K, V_h = X \cdot W_h^V, \quad (2)$$

where $W_h^Q, W_h^K, W_h^V \in \mathbb{R}^{d\times d_h}$ are the query, key, and value projection matrices for the h-th head, with dimension $d_h = d/H$. For each head $h \in \{1, \ldots, H\}$, the scaled dot-product attention is computed as:

$$A_h = \text{softmax}\left(\frac{Q_h K_h^\top}{\sqrt{d_h}}\right) \cdot V_h. \quad (3)$$

The outputs from all heads are then concatenated and transformed back to the original embedding space via a linear projection:

$$\text{MHA}(X) = \text{Concat}(A_1, A_2, \ldots, A_H) \cdot W^O, \quad (4)$$

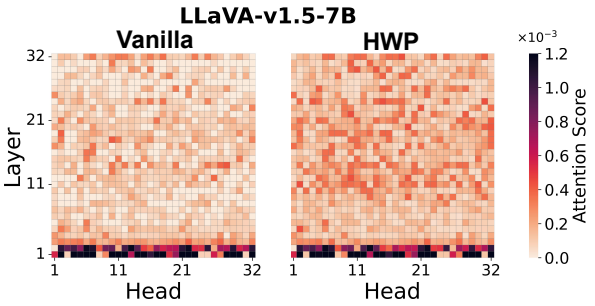

Figure 3: Visualization of mean attention scores from the final text token to all image tokens on each attention head. A significant proportion of baseline heads exhibit negligible attention to visual features. Conversely, our method fosters a more distributed allocation, where a broader set of heads actively attend to image tokens.

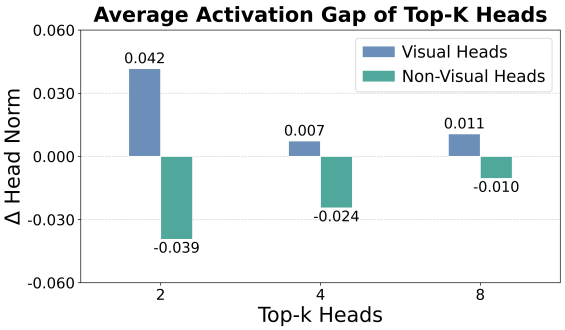

Figure 4: Average activation discrepancies by subtracting faithful samples to hallucinated samples. We report the difference in mean L_2 norm across the Top-k visual and non-visual attention heads, where $k \in \{2, 4, 8\}$. Hallucinated samples exhibit lower activation in visual heads and higher activation in non-visual heads.

where $W^O \in \mathbb{R}^{d \times d}$ is a learnable output projection matrix. This aggregation enables the model to integrate information from diverse attention subspaces, capturing complex contextual dependencies.

4 Method

4.1 Motivation

Despite the rapid evolution of VLMs (Wei et al., 2025; Team et al., 2025; Wang et al., 2024), these architectures remain susceptible to severe multimodal hallucinations, largely due to their insufficient utilization of visual information (Bi et al., 2025). Since the attention mechanism is the primary component for cross-modal integration in Transformer-based models, the degree to which attention heads engage with image tokens is a critical determinant of reliable vision-language grounding.

To investigate this, we visualize the average attention scores from the question-mark token in POPE (Li et al., 2023) hallucinatory queries (e.g., "<image> is there a dog?") to its corresponding image tokens using LLaVA-v1.5 (Liu et al., 2023). As illustrated in Figure 3, we notice a sparsity in visual attention that only a small subset of attention heads focus on image tokens, while the majority assign near-zero weights. This suggests that the model often operates under a state of visual neglect, relying instead on ingrained linguistic priors, thereby contributing to hallucinations.

Furthermore, we categorize attention heads based on visual focus, defining **visual heads** as the top-k heads with the highest mean attention scores toward image tokens, while **non-visual heads** as those with the lowest. Following (Kobayashi et al., 2020), we employ the L_2 norm of each head's output as a proxy for measuring its activation intensity. Our analysis in Figure 4 reveals a systemic imbalance: hallucinated samples (i.e., samples with incorrect responses) exhibit attenuated activation in visual heads but disproportionately stronger activation in non-visual heads compared to faithful samples (i.e., samples with correct responses).

These findings suggest that hallucinations are not merely a lack of attention, but are driven by a pathological activation shift toward heads that lack visual grounding. Consequently, we hypothesize that model reliability can be improved by distributing visual attention—essentially incentivizing a broader set of heads to allocate their representational capacity toward visual stimuli.

4.2 Continuous Noise Attention Perturbation

An intuitive approach to foster redundancy and encourage broader head participation in visual contexts is the application of dropout on attention heads. However, standard dropout—which acts as multiplicative noise sampled from a discrete binary distribution—often induces training instability due to the abrupt zeroing of entire head outputs. To address this issue, we propose an attention perturbation strategy that replaces discrete masking with **multiplicative uniform noise**. By introducing bounded continuous noise to attention outputs, our method ensures smoother gradient flow compared to binary dropout, while promoting a wider ensemble of heads to participate in processing visual features. This redistribution of attention capacity strengthens the model's focus on visual evidence and prevents visual information loss caused by abnormally activated attention heads, thereby mitigating potential hallucinations.

Individual attention heads are known to capture distinct functional aspects of a model's internal representations. These diverse features are concatenated and subsequently fused through the output projection matrix W^O. To formalize the relative contribution of each head, we introduce a vector of head-wise scaling factors $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_H]$, where H denotes the total number of attention heads. Using these coefficients, the multi-head attention output in Eq. (4) can be reformulated as:

$$\mathrm{MHA}(X) = \mathrm{Concat}(\lambda_1 A_1, \ldots, \lambda_H A_H) \cdot W^O \tag{5}$$

where A_h represents the output of the h-th head. In standard Transformer architectures, each attention head contributes equally to the final representation, corresponding to the special case where $\boldsymbol{\lambda} = \boldsymbol{I}$, signifying identical weighting across all heads.

To incentivize a broader set of heads to engage with image tokens, we introduce stochastic perturbations into the head-wise scaling factors. Specifically, for each attention head h, we modulate the coefficient by sampling a noise term from a uniform distribution as follows:

$$\boldsymbol{\lambda} = \mathbf{1} + a \cdot \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{U}(-1, 1)^d \tag{6}$$

where a controls the intensity of the noise perturbation. Notably, if $a = 1$ and $\boldsymbol{\epsilon}$ were sampled from a discrete Rademacher distribution (e.g., $\boldsymbol{\epsilon} \in \{-1, 1\}$), this formulation would degenerate into standard binary dropout.

Here, we opt for uniform noise over Gaussian or Rademacher alternatives. Compared to discrete Rademacher noise or unbounded Gaussian noise which may generate extreme activation outliers during training, continuous bounded uniform noise enables more stable optimization dynamics while providing sufficient variance to effectively regularize the attention mechanism.

4.3 Visual-Guided Training

To incentivize attention heads to prioritize visual evidence over linguistic shortcuts, we introduce a **visual-guided loss** that selectively targets **text tokens** strongly associated with visual input. Let $\mathcal{X} = \{x_i\}_{i=1}^N$ denote the input text token sequence and $\mathbf{I}_{\text{orig}} \in \mathbb{R}^{H \times W \times C}$ be the original image, where H, W, and C denote its height, width, and number of channels. To measure visual reliance, we construct a **random noise image** $\mathbf{I}_{\text{rand}} \in \mathbb{R}^{H \times W \times C}$, serving as an uninformative visual input baseline. Each pixel in $\mathbf{I}_{\text{rand}}$ is sampled independently from a discrete uniform distribution:

$$\mathbf{I}_{\text{rand}} \sim \mathcal{U}\{0, 1, \ldots, 255\}^{H \times W \times C}. \tag{7}$$

Let $z(x_i \mid \mathbf{I}) \in \mathbb{R}$ denote the scalar logit value corresponding to the specific text token x_i conditioned on image $\mathbf{I}$. We quantify the **visual sensitivity** of the i-th text token δ_i by calculating the difference between its logit values under the original image versus the random noise baseline:

$$\delta_i = z(x_i \mid \mathbf{I}_{\text{orig}}) - z(x_i \mid \mathbf{I}_{\text{rand}}). \tag{8}$$

A larger scalar value δ_i indicates stronger visual sensitivity, implying that removing visual information significantly alters the model's prediction for this token. To identify the most visually dependent tokens, we then define a threshold τ as the $(1-\beta)$-quantile of the sensitivity scores $\{\delta_i\}_{i=1}^N$. The set of **"visual tokens"**, denoted as $\mathcal{V}$, consists of the top $\beta\%$ of text tokens with the highest sensitivity:

$$\mathcal{V} = \{i \mid \delta_i > \tau\}. \tag{9}$$

The visual-guided loss is then formulated as loss restricted to this subset of visual tokens:

$$\mathcal{L}_{\text{visual}} = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \ell(\mathbf{z}_i, y_i), \tag{10}$$

where y_i is the ground-truth label and ℓ denotes the standard task loss (e.g., cross-entropy). This objective ensures that the optimization process emphasizes tokens strongly influenced by visual content, effectively guiding attention heads to capture visual features. To maintain the model's general-purpose capabilities, we employ LoRA for training. This strategy enhances visual grounding without disrupting the pre-trained knowledge base, preserving overall model versatility.

5 Experiments

5.1 Benchmarks

We evaluate our method across both generative and discriminative tasks (Appendix A) to assess its efficacy in mitigating multimodal hallucinations.

COCO-Caption To assess how well VLMs ground their captions in visual evidence and detect generative hallucinations, we apply the CHAIR metric (Rohrbach et al., 2018) on the MS COCO 2014 validation set (Lin et al., 2014). Model-generated captions are tokenized, mapped to MS

COCO object categories via a synonym list, and compared with ground-truth annotations to identify hallucinated objects. The sentence-level CHAIR_s quantifies the fraction of sentences containing at least one hallucinated object, while the instance-level CHAIR_i measures the proportion of generated objects that are hallucinated.

POPE Li et al. (2023) formulate hallucination detection as a binary classification task, where the model must determine the presence of a specific object or attribute in a given image. POPE evaluates object hallucinations in vision-language models using three distinct sampling strategies: Random, Popular, and Adversarial, enabling fine-grained evaluation of model robustness against object hallucinations of varying difficulty.

MME Fu et al. (2023) introduce a benchmark designed to evaluate both perception and cognition of multimodal large language models. Following Woo et al. (2025), we focus on the Existence, Count, Position, and Color sub-tasks to evaluate visual perception at both the object and attribute levels. Performance is evaluated using accuracy (Acc) and enhanced accuracy (Acc+), measuring correctness per question and correctness per image, respectively. We report a combined metric obtained by summing the two accuracies, consistent with established literature.

MMMU and MMMU-Pro Proposed by Yue et al. (2024), these benchmarks cover a wide range of open-ended and multiple-choice questions, designed to evaluate the general capabilities of VLMs across diverse domains. We report micro-averaged accuracy as the evaluation metric, which is calculated as accuracy weighted according to the number of examples of each category.

5.2 Experimental Setup

Base Models To evaluate the performance and scalability of our method, we conduct experiments across two representative families of VLMs. We include the widely-adopted LLaVA-v1.5 baseline (Liu et al., 2023) at the scales of 7B and 13B. Additionally, we evaluate our method on Qwen2.5-VL-7B-Instruct (Wang et al., 2024) (referred to as Qwen2.5-VL-7B in tables for brevity). This selection encompasses a diverse range of parameter magnitudes and architectures, facilitating a comprehensive assessment of our method's efficacy.

Base Model	Method	$\text{CHAIR}_s \downarrow$	$\text{CHAIR}_i \downarrow$	Avg. $\downarrow$
LLaVA-v1.5-7B	Vanilla	47.4	12.9	30.2
	Vanilla SFT	28.2	8.2	18.2
	VCD	45.6	11.4	28.5
	SPIN	26.8	7.2	17.0
	HWP (Ours)	**20.0**	**5.3**	**12.7**
LLaVA-v1.5-13B	Vanilla	42.8	10.6	26.7
	Vanilla SFT	28.6	8.3	18.5
	VCD	41.5	10.2	25.9
	SPIN	29.4	8.3	18.9
	HWP (Ours)	**18.2**	**6.2**	**12.2**
Qwen2.5-VL-7B	Vanilla	31.4	7.3	19.5
	Vanilla SFT	28.2	6.4	17.3
	VCD	31.2	7.4	19.3
	SPIN	30.6	7.1	18.9
	HWP (Ours)	**26.4**	**6.0**	**16.2**

Table 1: Results of object hallucination generation on the CHAIR metrics. Lower CHAIR_i and CHAIR_s indicate fewer hallucinations and better performance.

Implementation Details We adopt a 21.5k-sample vision-language instruction tuning dataset derived from Visual Genome (VG) (Krishna et al., 2017), as curated by Sarkar et al. (2024). To ensure computational efficiency and prevent catastrophic forgetting of the model's general capabilities, we employ LoRA (Hu et al., 2022) to train the VLMs for all experiments (denoted as the "SFT" setting). Unless otherwise specified, we configure LoRA with a rank $r = 256$ and a scaling factor $\alpha = 128$. The models are optimized with a learning rate of 2×10^{-5} for a single epoch. To maintain a controlled comparison across different architectures, the noise intensity a is fixed to 0.10, and set the image-token selection threshold to the 50th percentile. Throughout our experiments, we employ beam search with a beam width of 5. The maximum generation length is capped at 256 tokens. The prompt templates are presented in Appendix B.

Baselines In this section, we employ three baseline methods (Appendix C). First, we perform standard fine-tuning on the training dataset without introducing head noise or visual-guided loss (*Vanilla SFT*). In addition, we consider *VCD* (Leng et al., 2024), a traditional visual contrastive decoding approach that highlights tokens strongly associated with image tokens, and *SPIN* (Sarkar et al., 2025), which depresses the outputs of non-visual attention heads to mitigate language priors. Its effectiveness further suggests that attention head bias contributes to hallucinations.

Base Model	Method	POPE						MME			
		Advers.		Popular		Random		Exist.	Count	Pos.	Color
		Acc	F1	Acc	F1	Acc	F1	Score	Score	Score	Score
LLaVA-v1.5-7B	Vanilla	80.1	81.1	85.5	85.5	88.7	88.4	190.0	140.0	118.0	150.0
	Vanilla SFT	83.0	81.5	85.3	83.5	86.1	85.3	190.0	136.3	131.7	148.3
	VCD	81.2	81.7	85.7	85.6	88.6	88.2	185.0	145.0	126.7	145.0
	SPIN	82.7	82.5	86.2	85.7	88.9	88.3	**195.0**	146.7	118.3	155.0
	HWP (Ours)	**83.7**	**83.1**	**86.7**	**86.1**	**89.0**	**88.5**	190.0	**148.3**	**133.3**	**165.0**
LLaVA-v1.5-13B	Vanilla	81.6	82.5	84.8	85.2	88.8	88.6	185.0	128.3	101.7	155.0
	Vanilla SFT	82.6	82.8	86.4	86.2	89.6	88.8	180.0	138.3	110.0	160.0
	VCD	82.6	82.2	85.4	85.1	89.4	89.1	180.0	141.7	105.0	155.0
	SPIN	83.2	83.5	86.8	86.5	89.7	89.4	**190.0**	146.7	110.0	145.0
	HWP (Ours)	**83.7**	**84.0**	**87.3**	**87.2**	**90.4**	**90.0**	**190.0**	**155.0**	**125.0**	**175.0**
Qwen2.5-VL-7B	Vanilla	83.3	80.4	83.6	80.8	84.3	81.4	180.0	156.7	145.0	185.0
	Vanilla SFT	85.7	83.3	84.7	82.5	85.7	83.4	**185.0**	153.3	145.0	160.0
	VCD	83.7	82.3	84.5	83.2	84.7	83.8	**185.0**	156.7	150.0	185.0
	SPIN	85.2	83.4	85.7	84.4	85.9	84.8	**185.0**	153.3	150.0	175.0
	HWP (Ours)	**85.9**	**84.4**	**86.7**	**85.2**	**87.8**	**86.3**	**185.0**	**163.3**	**155.0**	**190.0**

Table 2: Results of hallucinatory discrimination on the POPE and MME benchmarks. The Accuracy and F1 are reported for POPE under Adversarial, Popular, and Random sampling settings. The total scores of summing Acc and Acc+ across four categories are reported for MME.

Strategy	Method	Exist.	Count	Pos.	Color
Greedy	Vanilla	**190.0**	125.0	95.0	160.0
	HWP (Ours)	**190.0**	**153.3**	**120.0**	**175.0**
Random	Vanilla	155.0	108.0	90.0	140.0
	HWP (Ours)	**185.0**	**125.0**	**101.7**	**156.7**

Table 3: Comparison of different decoding strategies on MME using LLaVA-v1.5-13B. Our method consistently increases base model performance.

Base Model	Method	MMMU	MMMU-Pro	Avg.
LLaVA-v1.5-7B	Vanilla	33.3	13.35	23.33
	HWP (Ours)	**33.5**	**13.53**	**23.52**
LLaVA-v1.5-13B	Vanilla	36.0	13.29	24.65
	HWP (Ours)	**36.2**	**14.86**	**25.53**
Qwen2.5-VL-7B	Vanilla	50.6	**29.65**	**40.13**
	HWP (Ours)	**51.1**	29.10	40.10

Table 4: Results of general-purpose capabilities on the MMMU and MMMU-Pro benchmarks. Our method does not downgrade the general capabilities of VLMs.

5.3 Result Analysis

Open-Ended Generation As demonstrated in Table 1, our method effectively anchors its descriptive synthesis in visual evidence, yielding a substantial reduction in generating object hallucinations across all model scales on COCO-Caption. Notably, on the LLaVA-v1.5-7B, our approach decreases CHAIR_s by **27.4** points and CHAIR_i by **7.6** points compared to the vanilla baseline.

Hallucination Discrimination We evaluate the performance of **HWP** on hallucination discriminative tasks using the POPE and MME benchmarks. As shown in Table 2, our method consistently enhances performance across various configurations and backbones. This demonstrates **HWP**'s ability in resisting object-presence biases, and fine-grained reasoning capabilities regarding attributes and spatial relations. Furthermore, as in Table 3, our method consistently improves performance on MME across other decoding strategies, including greedy and random, based on LLaVA-v1.5-13B.

General Capabilities To ensure that our attention perturbation does not compromise the core reasoning faculties of the models, we compare our method on the MMMU and MMMU-Pro benchmarks. As reported in Table 4, our method maintains competitive performance relative to the vanilla baselines without significant degradation, effectively enhancing visual grounding and mitigating hallucinations while successfully preserving the model's general-purpose capabilities. This underscores the practicality of our approach.

5.4 Ablation Studies

In this section, we conduct a series of ablation studies on LLaVA-v1.5-13B to systematically evaluate the contribution of each individual component in our framework. We first isolate the effects of attention perturbation and visual-guided loss to quantify their respective gains. Subsequently, we investigate the impact of different noise distributions and intensities on our approach, analyzing how these

Method	COCO-Caption		MME
	$\text{CHAIR}_s \downarrow$	$\text{CHAIR}_i \downarrow$	Score
Vanilla	42.8	10.6	570.0
+ SFT	28.6	8.3	588.3
+ SFT + AP	23.0	7.6	625.0
+ SFT + VL	22.0	7.4	615.0
+ SFT + AP + VL	**18.2**	**6.2**	**645.0**

Table 5: Ablation studies of architectural components, evaluated on the COCO-Caption and MME.

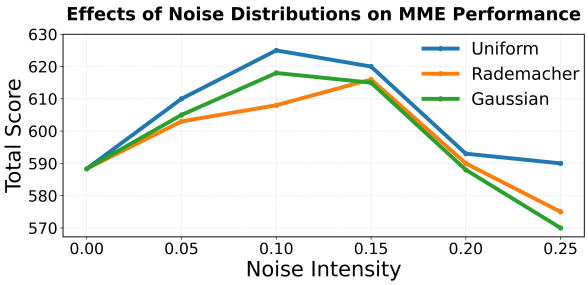

Figure 5: Performance comparison of different noise distributions on the MME benchmark. Varying noise intensities are sampled on each distribution.

design choices influence model performance. Finally, we examine the effect of varying the visual token selection ratio for loss computation.

Architectural Components As shown in Table 5, supervised fine-tuning with LoRA yields a performance gain of 14.2 and 2.3 points on CHAIR_s and CHAIR_i, but remains insufficient for fully addressing multimodal hallucinations. In contrast, incorporating head noise and visual-guided loss leads to more substantial improvements across all benchmarks, demonstrating our effectiveness.

Noise Distributions and Intensities We compare the efficacy of various stochastic perturbations in Table 6. For each noise distribution, we report the optimal performance identified through a grid search over noise intensities $a \in [0.05, 0.25]$. For dropout, we restrict dropout ratios $p \in [0.01, 0.15]$, as performance on COCO-Caption deteriorates sharply when $p \geq 0.15$. Peak performance for dropout is achieved at $p = 0.08$. Among the evaluated perturbations, **uniform noise** yields the best performance as compared to standard Gaussian and Rademacher alternatives, proving that it stands as an effective choice for promoting a balanced redistribution of visual attention across heads.

Using uniform noise, we further analyze the sensitivity to noise intensity without visual-guided loss in Figure 5. We observe a non-monotonic MME performance trend: results initially improve but begin to decline beyond a threshold of $a = 0.10$. This behavior indicates that while moderate stochasticity effectively breaks the model's over-reliance on a few dominant attention heads, excessive noise levels inevitably hinder the model's ability to learn consistent cross-modal mappings.

Method	$\text{CHAIR}_s \downarrow$	$\text{CHAIR}_i \downarrow$
Vanilla	42.8	10.6
+ Dropout	25.6	7.9
+ Rademacher Noise	24.0	7.9
+ Gaussian Noise	23.4	7.7
+ Uniform Noise	**23.0**	**7.4**

Table 6: Ablation studies of attention noise perturbation type, evaluated on the COCO-Caption.

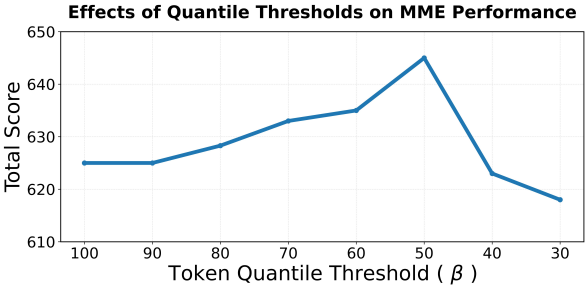

Figure 6: Impact of visual token selection ratio (β) on MME performance. The value $\beta = 100$ denotes the all-token standard training baseline, while a smaller threshold represents selecting fewer visual tokens of stronger correlation with visual content for training.

Visual Token Selection Ratio We investigate the impact of the selection ratio of visual tokens in Figure 6. As the quantile threshold decreases and filters out more tokens, MME performance exhibits an initial upward trend before eventually declining after the 50th quantile. This finding offers a valuable insight for the community: during VLM training, selectively masking textual tokens with weak visual correlations can encourage the model to prioritize visual evidence and reduce its reliance on linguistic shortcuts, thereby effectively mitigating the generation of hallucinations.

6 Conclusion

In this work, we analyze multi-head attention in VLMs and identify a key bottleneck: most heads are weakly sensitive to visual information. We further show that hallucinations arise from specific activation patterns where visual heads are underactivated. To address this issue, we propose a training strategy that injects uniform noise into attention head outputs and applies a visual-guided loss on

vision-sensitive tokens. By disrupting dominant activation patterns and emphasizing visual relevance, our method enhances visual awareness across attention heads. Extensive experiments across benchmarks and model scales show that our approach consistently reduces hallucinations while preserving general-purpose capabilities.

Limitations

We note that our training introduces additional computation to identify tokens strongly associated with visual information, which results in increased training latency. We do not investigate the underlying mechanisms behind why certain attention heads tend to underweight visual information. While understanding these mechanisms may provide further insights into hallucination behavior and help guide future improvements, this is left for future work.

References

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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4135–4144.

Cong Chen, Mingyu Liu, Chenchen Jing, Yizhou Zhou, Fengyun Rao, Hao Chen, Bo Zhang, and Chunhua Shen. 2025a. Perturbollava: Reducing multimodal hallucinations with perturbative visual training. *arXiv preprint arXiv:2503.06486*.

Kaiyuan Chen, Shuangyu Xie, Zehan Ma, Pannag R Sanketi, and Ken Goldberg. 2025b. Robo2vlm: Visual question answering from large-scale in-the-wild robot manipulation datasets. *arXiv preprint arXiv:2505.15517*.

Jingyuan Deng and Yujiu Yang. 2025. Maskcd: Mitigating lvlm hallucinations by image head masked contrastive decoding. *arXiv preprint arXiv:2510.02790*.

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

Aarti Ghatkesar, Uddeshya Upadhyay, and Ganesh Venkatesh. 2025. Looking beyond language priors: Enhancing visual comprehension and attention in multimodal models. *arXiv preprint arXiv:2505.05626*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. 2025. See what you are told: Visual attention sink in large multimodal models. *arXiv preprint arXiv:2503.03321*.

Jinyeong Kim, Seil Kang, Jiwoo Park, Junhyeok Kim, and Seong Jae Hwang. 2025. Interpreting attention heads for image-to-text information flow in large vision-language models. *arXiv preprint arXiv:2509.17588*.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, and 1 others. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Yi-Lun Lee, Yi-Hsuan Tsai, and Wei-Chen Chiu. 2024. Delve into visual contrastive decoding for hallucination mitigation of large vision-language models. *arXiv preprint arXiv:2412.06775*.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Yinan Li and Fang Liu. 2016. Whiteout: Gaussian adaptive noise regularization in deep neural networks. *arXiv preprint arXiv:1612.01490*.

Xiaoyu Liang, Jiayuan Yu, Lianrui Mu, Jiedong Zhuang, Jiaqi Hu, Yuchen Yang, Jiangnan Ye, Lu Lu, Jian Chen, and Haoji Hu. 2024. Mitigating hallucination in visual-language models via re-balancing contrastive decoding. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 482–496. Springer.

Min Lin, Xiwen Liang, Bingqian Lin, Liu Jingzhi, Zijian Jiao, Kehan Li, Yuhan Ma, Yuecheng Liu, Shen Zhao, Yuzheng Zhuang, and 1 others. 2025. Echovla: Robotic vision-language-action model with synergistic declarative memory for mobile manipulation. *arXiv preprint arXiv:2511.18112*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Yang Liu, Xinshuai Song, Kaixuan Jiang, Weixing Chen, Jingzhou Luo, Guanbin Li, and Liang Lin. 2024. Meia: Multimodal embodied perception and interaction in unknown environments. *arXiv preprint arXiv:2402.00290*.

Eric Nalisnick, Anima Anandkumar, and Padhraic Smyth. 2015. A scale mixture perspective of multiplicative noise in neural networks. *arXiv preprint arXiv:1506.03208*.

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*, pages 4035–4045.

Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Ö Arık, and Tomas Pfister. 2024. Mitigating object hallucination in mllms via data-augmented phrase-level alignment. *arXiv preprint arXiv:2405.18654*.

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*, pages 12492–12511.

Xu Shen, Xinmei Tian, Tongliang Liu, Fang Xu, and Dacheng Tao. 2017. Continuous dropout. *IEEE transactions on neural networks and learning systems*, 29(9):3926–3937.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Feilong Tang, Chengzhi Liu, Zhongxing Xu, Ming Hu, Zile Huang, Haochen Xue, Ziyang Chen, Zelin Peng, Zhiwei Yang, Sijin Zhou, and 1 others. 2025. Seeing far and clearly: Mitigating hallucinations in mllms with attention causal decoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26147–26159.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jiahui Wang, Zuyan Liu, Yongming Rao, and Jiwen Lu. 2025. Sparsemm: Head sparsity emerges from visual concept responses in mllms. *arXiv preprint arXiv:2506.05344*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Haoran Wei, Yaofeng Sun, and Yukun Li. 2025. Deepseek-ocr: Contexts optical compression. *arXiv preprint arXiv:2510.18234*.

Sangmin Woo, Donguk Kim, Jaehyuk Jang, Yubin Choi, and Changick Kim. 2025. Don't miss the forest for the trees: Attentional vision calibration for large vision language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1927–1951.

Jingkang Yang, Yuhao Dong, Shuai Liu, Bo Li, Ziyue Wang, Haoran Tan, Chencheng Jiang, Jiamu Kang, Yuanhan Zhang, Kaiyang Zhou, and 1 others. 2024a. Octopus: Embodied vision-language programmer from environmental feedback. In *European conference on computer vision*, pages 20–38. Springer.

Yi Yang, Jiaxuan Sun, Siqi Kou, Yihan Wang, and Zhijie Deng. 2025. Lohovla: A unified vision-language-action model for long-horizon embodied tasks. *arXiv preprint arXiv:2506.00411*.

Yijun Yang, Tianyi Zhou, Kanxue Li, Dapeng Tao, Lusong Li, Li Shen, Xiaodong He, Jing Jiang, and Yuhui Shi. 2024b. Embodied multi-modal agent trained by an llm from a parallel textworld. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26275–26285.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

A Benchmark Details

COCO-Caption CHAIR (Rohrbach et al., 2018) evaluates hallucination at both the instance and sentence levels. The instance-level metric, CHAIR_i, measures the fraction of object mentions generated by the model that do not appear in the ground-truth annotations:

$$\text{CHAIR}_i = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all objects mentioned}\}|}. \quad (11)$$

In contrast, the sentence-level metric, CHAIR_s, measures the proportion of sentences that contain at least one hallucinated object:

$$\text{CHAIR}_s = \frac{|\{\text{sentences with hallucinated objects}\}|}{|\{\text{all sentences}\}|}. \quad (12)$$

Together, CHAIR_i and CHAIR_s provide complementary perspectives on hallucination behavior, capturing both the overall frequency of hallucinated object mentions and their distribution across sentences.

POPE The POPE dataset (Li et al., 2023) is designed to evaluate object hallucination in vision-language models under controlled negative sampling. It provides three distinct configurations, which progressively increase in difficulty:

- **Random:** Negative objects are sampled uniformly from categories that are not present in the image. This setting provides a relatively easy baseline, as the sampled objects are generally unrelated to the visual content.
- **Popular:** Negative objects are selected from high-frequency categories in the dataset that are absent from the image. This setting challenges the model's reliance on language priors, as it may be biased toward predicting common objects even when they are not visually present.
- **Adversarial:** Negative objects are chosen to be semantically related to objects present in the image but are themselves absent. This setting presents the hardest challenge, testing the model's ability to avoid hallucinating contextually plausible but visually unsupported objects.

For each setting, the POPE test set contains $3,000$ samples. We report both accuracy and F1 score.

MME The coarse-grained recognition subset of the MME benchmark (Fu et al., 2023) evaluates VLMs on their ability to understand general object-level properties within an image. This subset consists of four perception-oriented tasks:

- **Existence:** Determine whether a particular object is present in the image.
- **Position:** Assess the model's understanding of spatial relationships by identifying the location of an object within the image, such as top-left, center, or relative to other objects.
- **Color:** Evaluate the model's recognition of object attributes by predicting the color of a specified object.
- **Count:** Measure the model's ability to enumerate instances of a specific object category within an image.

Images are sampled from COCO (Lin et al., 2014), but the instruction-answer pairs are manually constructed. Even if the model has encountered these COCO images during pretraining, these specific instruction-answer pairs are not included in its training set. For each subtask (existence, count, color, position), 30 images are selected, and for each image, two instruction-answer pairs are provided, resulting in 60 pairs per subtask. This design ensures that models must understand the instructions and reason over visual content to infer correct answers, rather than relying on memorization.

B Prompt Templates

LLaVA-v1.5-7B / LLaVA-v1.5-13B

```
USER: <image> Please describe this image in detail.
ASSISTANT:
```

Qwen2.5-VL-7B

```
<|im_start|>system
You are a helpfulassistant.<|im_end|>
<|im_start|>user
Please describe this image in detail.
<|vision_start|><|image_pad|><|vision_end|><|im_end|>
<|im_start|>assistant
```

Figure 7: The chat template used in our experiments. Tokens highlighted in red indicate the placeholder positions for image tokens in multimodal inputs.

C Baseline Details

VCD VCD (Leng et al., 2024) highlights tokens that are strongly associated with visual information through contrastive decoding. It involves three hyperparameters. The first is the noise level applied to the input image, which is defined by the diffusion step: a larger step results in more severe corruption of the original visual information. We set the noise step to 900, where the maximum diffusion step is $1,000$. The second hyperparameter is the contrastive strength α, which we set to 1.0. The third hyperparameter is β, which controls the strictness of the plausibility constraint: larger values of β retain only high-confidence tokens and thus avoid indiscriminate penalization of linguistically plausible outputs. We set β to 0.1.

SPIN SPIN (Sarkar et al., 2025) suppresses the weights of non-visual attention heads to reduce the model's reliance on language priors. The method involves two hyperparameters: α, which controls the suppression strength (with $\alpha = 0$ indicating that the output of the suppressed heads is completely removed), and γ, which determines the proportion of suppressed heads within each layer. Following the recommended settings in SPIN, we adopt task- and model-specific configurations. For the **COCO Caption** task, SPIN is applied as follows: layers 1–32 for LLaVA-v1.5-7B with $\alpha = 0.08$ and $\gamma = 0.05$; layers 1–16 for LLaVA-v1.5-13B with $\alpha = 0.0$ and $\gamma = 0.10$; and layers 1–20 for Qwen-2.5-VL with $\alpha = 0.08$ and $\gamma = 0.30$. For **discriminative tasks** (POPE and MME), SPIN is applied as follows: layers 1–32 for LLaVA-v1.5-7B with $\alpha = 0.1$ and $\gamma = 0.20$; layers 1–20 for LLaVA-v1.5-13B with $\alpha = 0.0$ and $\gamma = 0.35$; and layers 1–20 for Qwen-2.5-VL with $\alpha = 0.08$ and $\gamma = 0.30$.