Not All Tokens and Heads Are Equally Important: Dual-Level Attention Intervention for Hallucination Mitigation

Lexiang Tang[1], Xianwei Zhuang[1], Bang Yang[1,2], Zhiyuan Hu[1], Hongxiang Li[3], Lu Ma[4], Jinghan Ru[1], Yuexian Zou[1,2*]

[1]Guangdong Provincial Key Laboratory of Ultra High Definition Immersive Media Technology, Shenzhen Graduate School, Peking University
[2]Pengcheng Laboratory
[3] The Hong Kong University of Science and Technology
[4]Peking University
{tanglexiang, xwzhuang, huzhiyuan, lihongxiang, maluqaq, rujinghan}@stu.pku.edu.cn {yangbang,zouyx}@pku.edu.cn

Abstract

Large vision-language models (LVLMs) have demonstrated impressive capabilities across diverse multimodal tasks, yet they remain highly susceptible to visual hallucinations (VH), often producing confident but inaccurate descriptions of visual content. Building on the insight that not all tokens and attention heads contribute equally to VH mitigation, we introduce VisFlow, a lightweight and training-free framework that alleviates hallucinations by directly modulating attention patterns during inference. To address two primary challenges of VH, namely insufficient visual attention and the dominance of language priors, we identify three problematic attention behaviors in LVLMs: (1) disproportionate allocation of attention to uninformative or trailing visual tokens, (2) over-dependence on the previously generated token, and (3) excessive fixation on system prompts that hinders multimodal integration. To overcome these issues, VisFlow introduces a dual-level Attention Intervention, consisting of Token-level Attention Intervention (TAI), which reinforces attention to salient visual regions, and Head-level Attention Intervention (HAI), which suppresses undue focus on system prompts and adjacent text tokens. Together, these interventions strengthen visual alignment while reducing linguistic bias. Extensive experiments across diverse models and benchmarks demonstrate that VisFlow effectively mitigates hallucinations with minimal computational overhead.

Introduction

Built upon the rapid progress of Large Language Models (LLMs) (Yang et al. 2025a; Team 2024; Touvron et al. 2023; Chiang, Li et al. 2023; Ru et al. 2025), Large Vision-Language Models (LVLMs) (Bai et al. 2025; Wang et al. 2024a; Chen et al. 2023b; Liu et al. 2023; Ye et al. 2024; Bai et al. 2023; Chen et al. 2024b; Li et al. 2023a; Chen et al. 2023a; Zhuang et al. 2025a) have achieved strong performance across a wide range of multimodal understanding and generation tasks. An LVLM typically handles four types of tokens: (1) system prompts configuring model behavior, (2) visual tokens encoding image content, (3) instruction tokens representing user queries, and (4) response tokens as output text. Despite their capabilities, LVLMs often generate outputs misaligned with the visual input—a phenomenon known as visual hallucination (VH) (Liu et al. 2024b; Leng et al. 2024; Huang et al. 2024), which poses risks in real-world and safety-critical applications.

Existing VH mitigation strategies fall into three categories: (1) Instruction tuning with hallucination-aware datasets (Jiang et al. 2024; Sarkar et al. 2024; Chen et al. 2025; Yang et al. 2025c), which improves grounding but requires costly retraining; (2) Auxiliary modules such as reranking or hallucination detection (Manakul, Liusie, and Gales 2023; Yin et al. 2024; Chen et al. 2024c), which introduce additional latency and complexity; and (3) Decoding-time interventions (Huang et al. 2024; Wang et al. 2024b; Fan et al. 2025; Liu et al. 2024c; Zhuang et al. 2025b), which are more efficient but often depend on contrastive decoding or external grounding tools, limiting scalability. In this work, we present VisFlow, a training-free framework that mitigates VH at inference by directly modulating attention patterns within the LVLM decoder. Building on prior analyses (Liu et al. 2024b; Yin, Si, and Wang 2025), we focus on two primary contributing factors: (1) insufficient attention to visual information, and (2) over-reliance on language priors.

Addressing the first issue, PAI (Liu, Zheng, and Chen 2024) enhances attention weights equally across all visual tokens, while VAR (Kang et al. 2025) introduces the concept of the visual sink token and reallocates redundant attention from the BOS and visual sink tokens to other visual tokens to improve visual alignment. However, these approaches lack precise targeting and may inadvertently reinforce cross-modal attention errors introduced by Rotary Positional Embedding (RoPE) (Su et al. 2024). CCA-LLaVA (Xing et al., 2024) shows that RoPE causes text tokens to over-attend to nearby visual ones.

To overcome these limitations, we identified a critical insight: not all visual tokens are equally important for mitigating hallucination. Leveraging this, we propose Token-Level Attention Intervention (TAI), which selectively enhances attention to these crucial visual tokens while also correcting for RoPE-induced bias. To our knowledge, we are the first to identify critical visual tokens such as the visual sink and salient tokens based on intra-modal attention. In-

*Corresponding author.

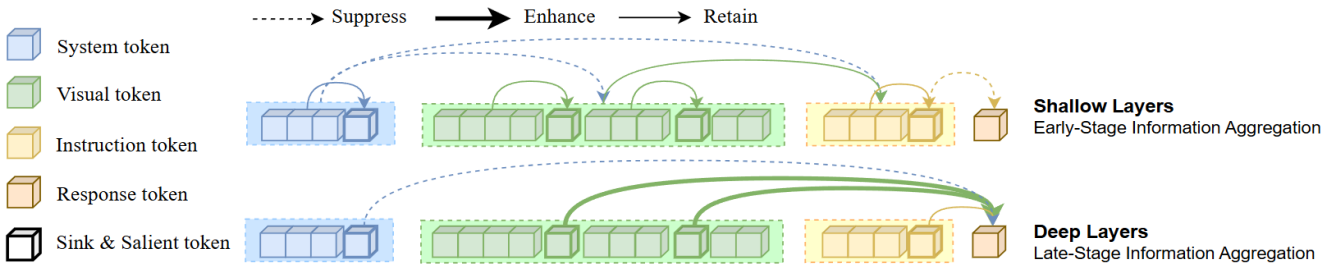


Figure 1: Illustration of token-level attention distribution in VisFlow. We allocate stronger attention to visual salient token while avoiding over-attention to system prompts and prior text tokens, resulting in more balanced cross-modal alignment.

spired by prior work (Darcet et al. 2023; Wang et al. 2023), which suggests that the visual sink token aggregates global semantic information from preceding tokens, we hypothesize that suppressing attention to the sink token may impair the model's perception of global semantics. We validate this hypothesis through targeted intervention experiments on the POPE (Li et al. 2023b), specifically focusing on the sink token. Unlike VAR, which redistributes attention away from it, TAI amplifies attention to the sink token on POPE.

Addressing the second issue, contrastive decoding paradigms such as VCD (Leng et al. 2024) and ICD (Wang et al. 2024b) mitigate hallucination by introducing perturbations to the visual or textual inputs, thereby increasing the uncertainty of model outputs and producing contrastive distributions dominated by linguistic priors. These contrastive distributions are then subtracted from the original ones to suppress hallucinations. However, generating these contrastive distributions requires multiple forward passes, leading to significant inference overhead.

To overcome this, we introduce Head-Level Attention Intervention (HAI), a method built on a critical insight: not all heads are equally important for mitigating hallucination. Our approach is the first to explicitly identify text attention heads correlated with linguistic priors and reduce their abnormal attention on text tokens. Since HAI directly modifies the original attention distribution, it provides a clear efficiency advantage over contrastive decoding methods (Leng et al. 2024; Park et al. 2025; Huo et al. 2024; Wang et al. 2024b; An et al. 2025). Our method avoids their computational burden.

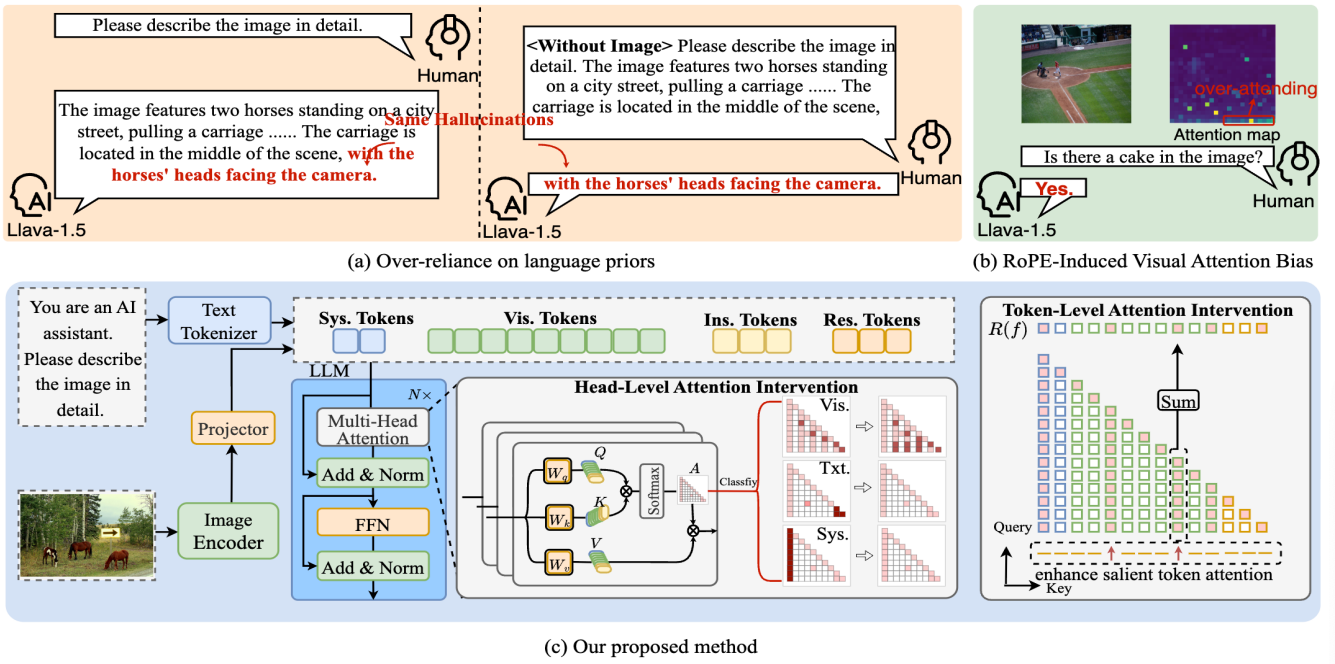
Our main contributions are as follows. (1) We analyze VH in LVLMs through the lens of attention and information flow, identifying different types of visual tokens and attention heads associated with hallucination. (2) We introduce VisFlow, a training-free and efficient inference-time framework incorporating TAI and HAI to enhance visual alignment and suppress hallucinations. (3) Comprehensive experiments demonstrate that VisFlow outperforms existing methods in both effectiveness and efficiency.

Related Work

Large Vision-Language Model (LVLM) LVLMs combine large language models (LLMs) with visual encoders to handle multimodal inputs. Early approaches, such as the LLaVA series (Liu et al. 2024a, 2023), align visual features with LLMs via linear projections and enhance performance through instruction tuning. Other works like BLIP-2 (Li et al. 2023a), MiniGPT-4 (Chen et al. 2023a), and InstructBLIP (Liu et al. 2023) introduce query transformers (e.g., Q-former (Li et al. 2023a)) to extract instruction-aware visual features for improved efficiency. Recent models such as Qwen2.5-VL (Bai et al. 2025, 2023), mPLUG-Owl2 (Ye et al. 2024), and InternVL2.5 (Chen et al. 2024a,b) further optimize architectures, training, and data pipelines.

Visual Hallucination Mitigation Efforts to mitigate visual hallucination (VH) in LVLMs fall into three categories: (1) Instruction tuning (Gunjal, Yin, and Bas 2024; Jiang et al. 2024), which improves grounding but requires task-specific data and costly retraining; (2) Auxiliary analysis (Manakul, Liusie, and Gales 2023; Chen et al. 2024c; Yin et al. 2024), which adds inference-time modules at the cost of latency; and (3) Decoding-time interventions (Huang et al. 2024; Wang et al. 2024b; Fan et al. 2025; Liu et al. 2024c; Zhuang et al. 2025b), which intervene during generation and are more efficient than retraining or auxiliary modules, but often rely on contrastive decoding or external grounding tools, increasing inference latency. Recent attention-based interventions (Yang et al. 2025b; Kang et al. 2025; Yin, Si, and Wang 2025) directly modify attention distributions during inference, avoiding extra decoding steps, but often lack fine-grained targeting, focusing either on attention heads (Yang et al. 2025b; Yin, Si, and Wang 2025) or token importance (Zou et al. 2024; Kang et al. 2025). Our method VisFlow integrates TAI and HAI to jointly identify salient tokens and critical heads, enabling precise attention correction with minimal computational overhead.

Information Flow in MLLMs Zhang et al. (Zhang et al. 2024) provide an empirical visualization of cross-modal information flow in the decoder of LVLMs, outlining a three-stage process from global visual feature integration to final output generation. However, their analysis overlooks interactions among visual tokens. We address this gap by analyzing attention between visual tokens. Additionally, CCA-LLaVA (Xing et al. 2024) attributes VH to disrupted token-level information flow, particularly the long-range decay introduced by RoPE. Our proposed TAI module mitigates RoPE-induced bias. Building on this insight, as shown in Figure 1 and 3, we analyze the token-level information flow and attention distribution in LVLMs.



(a) Over-reliance on language priors

(b) RoPE-Induced Visual Attention Bias

(c) Our proposed method

Figure 2: Overview of visual hallucination causes and our solution. (a) Linguistic Over-reliance: hallucinations caused by excessive dependence on language priors; (b) RoPE-induced Attention Bias: attention skewed toward image tokens near text tokens; (c) Our Method: mitigates these issues via Token-level Attention Intervention (TAI) to enhance focus on salient visual cues, and Head-level Attention Intervention (HAI) to suppress over-attention to system and nearby text tokens.

Method

This section analyzes token-level information flow and attention distribution in MLLMs, and introduces our methods.

Information Flow in Multimodal Tokens

To analyze how MLLMs utilize visual information and why they may overly rely on language priors, we adopt the saliency technique (Simonyan, Vedaldi, and Zisserman 2013), to highlight critical token interactions within the attention mechanism. The saliency score is computed by taking the Hadamard product of the attention scores A and their gradients as follows:

$$I_l = \left| \sum_h A_{h,l} \odot \frac{\partial \mathcal{L}(x)}{\partial A_{h,l}} \right|, \tag{1}$$

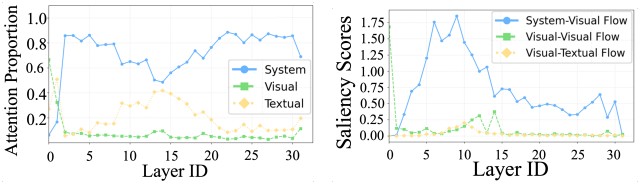
where $A_{h,l}$ is the attention matrix from the h-th head in the l-th layer, and $\mathcal{L}(x)$ denotes the task loss. The saliency matrix I_l aggregates all heads to reflect the contribution of token j to token i in layer l.

To quantify directional information flow between token groups (e.g., system, visual, textual), we define:

$$S_{ab} = \frac{\sum_{(i,j)\in C_{ab}} I(i,j)}{|C_{ab}|}, \tag{2}$$

$$C_{ab} = \{(i,j) : i \in \mathcal{A},\ j \in \mathcal{B}\}, \tag{3}$$

where $I(i,j)$ measures information flow from token j to token i, and C_{ab} is the set of all such directed token pairs. For example, S_{sv} measures system-to-visual flow, S_{vv} captures intra-visual flow (optionally constrained by $i \geq j$), and S_{vt} quantifies visual-to-text transfer. Figures 3a and 3b consistently show biased attention patterns: insufficient attention and information flow on visual tokens and abnormally high focus on system prompts, suggesting impaired visual alignment.



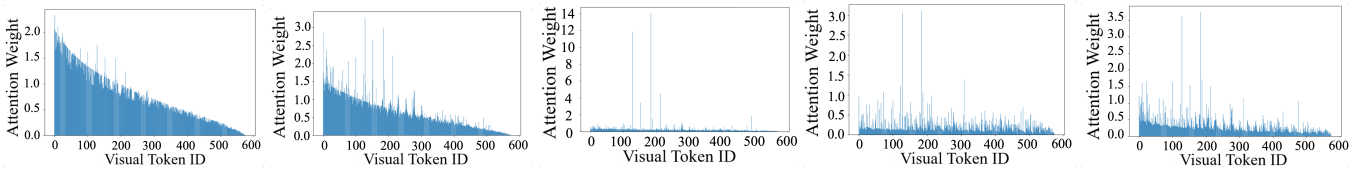
(a) Attention distribution.

(b) Salience scores.

Figure 3: Layer-wise token attention and information flow analysis on 500 MSCOCO samples from the POPE.

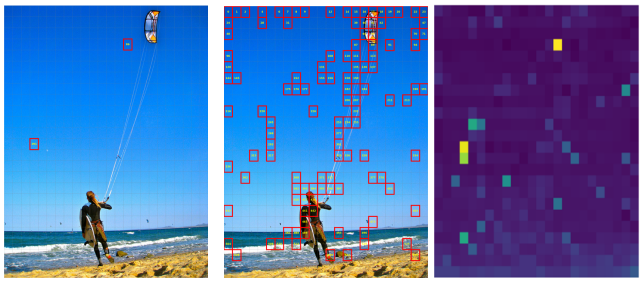
Token-Level Attention Intervention

Visual Sink and Salient Token Identification To address misaligned visual attention, partly due to RoPE-induced position bias, we identify visual-salient tokens by analyzing intra-visual attention patterns. We define the reception score $R(j)$ of a visual token $j \in \mathcal{I}_{\text{vis}}$ as the total attention it receives from other visual tokens across all heads in a decoder layer:



(a) Layer 0. (b) Layer 1. (c) Layer 2. (d) Layer 15. (e) Layer 31.

Figure 4: Visualization of intra-visual reception score $R(j)$ among visual tokens across decoder layers.



(a) Sink token (b) Salient token (c) Attention map

Figure 5: Visualization of visual attention in LLaVA. (a) Visual Sink Token: tokens that absorb much attention but lacks semantic contribution; (b) Visual Salient Token: tokens that align with meaningful visual regions critical for grounding; (c) Attention distribution over visual tokens.

$$\mathrm{R}(j) = \frac{1}{H} \sum_{h=1}^{H} \sum_{i \in \mathcal{I}_{\text{vis}} \setminus \{j\}} \mathbf{A}_{\ell}^{(h)}[i, j], \tag{4}$$

where $\mathbf{A}_{\ell}^{(h)}$ is the attention matrix of head h in layer ℓ, and H is the number of heads. Figure 4 shows the distribution of $R(j)$ across layers, revealing uneven attention among visual tokens. As shown in Figure 4a and Figure 4b, early layers exhibit globally distributed interactions among visual tokens, with clear bias introduced by causal attention. However, starting from layer 2, distinct visual tokens with significantly higher $R(j)$ begin to emerge. These tokens correspond to the visually salient or sink tokens we aim to identify. To extract them, we select tokens whose $R(j)$ exceed a fraction τ of the maximum score:

$$\mathcal{I}_{\text{thres}}(\tau) = \left\{ j \in \mathcal{I}_{\text{vis}} \,\middle|\, \mathrm{R}(j) > \tau \cdot \max_{k \in \mathcal{I}_{\text{vis}}} \mathrm{R}(k) \right\}, \tag{5}$$

where a small τ (e.g., $\frac{1}{20}$) selects salient tokens, while a larger τ (e.g., $\frac{1}{2}$) indicates sink tokens.

As shown in Figure 5, we visualize the attention distribution of the last text token to all visual tokens in Layer 15 of LLaVA-1.5-7B, along with the identified sink and salient tokens for a given image, highlighting any misalignments between the attention map and true visual relevance.

Enhancing Intra-Visual Salient Token Attention To improve visual alignment and suppress attention bias introduced by RoPE encoding, we modify the attention weights from instruction tokens to visual tokens at each decoder layer ℓ and head h. Specifically, we amplify attention toward salient regions and attenuate it for semantically meaningless sink tokens. The adjusted attention weights are defined as:

$$A_{i,j}^{\ell,h} = \begin{cases} k \cdot A_{i,j}^{\ell,h}, & \text{if } i \in \mathcal{I}_{\text{txt}},\ j \in \mathcal{I}_{\text{salient}}^{\ell} \\ \delta \cdot A_{i,j}^{\ell,h}, & \text{if } i \in \mathcal{I}_{\text{txt}},\ j \in \mathcal{I}_{\text{sink}}^{\ell} \end{cases} \tag{6}$$

where $k > 1$ is a scaling factor that enhances attention to salient visual tokens, and $\delta < 1$ is a decay factor that suppresses attention to sink tokens.

To ensure the attention weights sum to 1, the modified weights are re-normalized as follows:

$$A_{i,j}^{\ell,h} = \frac{A_{i,j}^{\ell,h}}{\sum_{j} A_{i,j}^{\ell,h}}, \quad \text{if } i \in \mathcal{I}_{\text{txt}}. \tag{7}$$

This strategy encourages the model to attend more effectively to informative visual evidence while reducing focus on irrelevant regions.

Head-Level Attention Intervention

Attention Head Type Identification Prior studies (Sun et al. 2024; Tang et al. 2025; Michel, Levy, and Neubig 2019; Yang et al. 2025b) have shown that attention heads in large models often assume specialized functions. To investigate how different token types shape attention in decoder, we categorize heads into three types: (1) Visual-sensitive, focusing on visual tokens; (2) System-dominant, attending mainly to system prompts; (3) Text-dominant, aligned with instruction and response tokens.

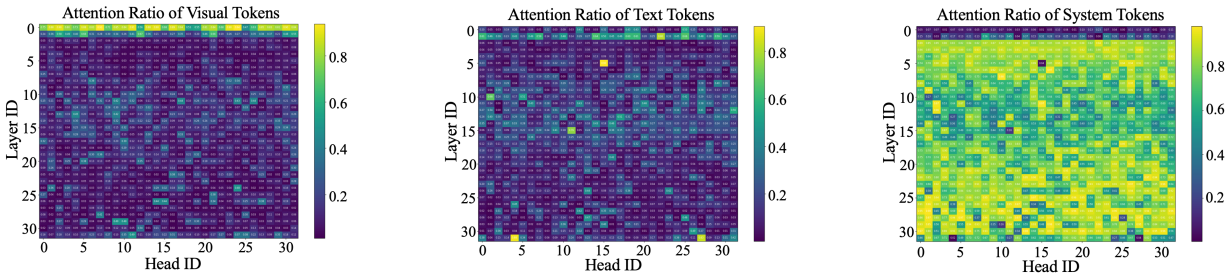
For each head h in decoder layer ℓ, we measure its attention to visual tokens as:

$$A_{\text{vis}}^{\ell,h} = \sum_{i \in \mathcal{I}_{\text{txt}}} \sum_{j \in \mathcal{I}_{\text{vis}}} \mathbf{A}^{\ell,h}[i, j], \tag{8}$$

where $\mathcal{I}_{\text{txt}}$ denotes the set of instruction tokens during the prefill stage and the generated tokens during the decode stage. A head is considered visual-sensitive if its visual attention exceeds a significance threshold:

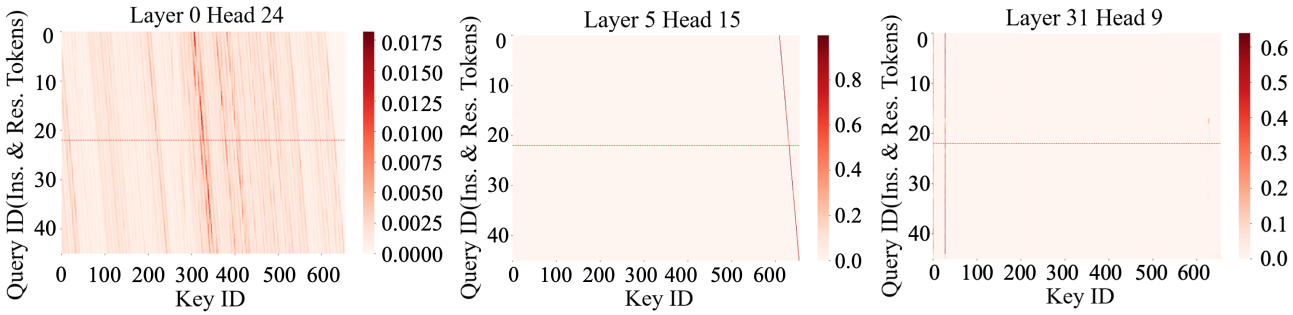
$$\mathcal{H}_{\text{vis}}^{\ell} = \left\{ h \,\middle|\, A_{\text{vis}}^{\ell,h} > \mu + \lambda_{\text{vis}} \cdot \sigma \right\}, \tag{9}$$

where μ and σ are the mean and standard deviation of $A_{\text{vis}}^{\ell,h}$ across all heads in layer ℓ, and λ_{vis} is a tunable hyperparameter controlling sensitivity. This selection identifies heads that exhibit unusually strong attention toward visual input.



(a) Visual attention (b) Text attention (c) System attention

Figure 6: Layer-wise heatmaps of attention weights across all heads.



(a) Visual head (b) Text head (c) System head

Figure 7: Attention maps from representative heads across modalities. Attention is visualized from instruction tokens (prefill stage) and response tokens (decode stage) to all key tokens. Red line separates the two phases.

For each attention head h in layer ℓ, we compute the total attention directed from instruction tokens to a target token set $\mathcal{C}$, defined as:

$$A_{\mathcal{C}}^{\ell,h} = \begin{cases} \sum_{i \in \mathcal{I}_{\text{txt}}} \sum_{j \in \mathcal{I}_{\text{txt}}} \mathbf{A}^{\ell,h}[i,j], & \text{if } \mathcal{C} = \text{txt} \\ \sum_{i \in \mathcal{I}_{\text{txt}}} \sum_{j \in \mathcal{I}_{\text{sys}}} \mathbf{A}^{\ell,h}[i,j], & \text{if } \mathcal{C} = \text{sys} \end{cases} \tag{10}$$

where $\mathcal{C} \in \{\text{txt}, \text{sys}\}$ denotes the token type of interest (i.e., text or system prompt tokens). Attention heads with dominant focus on $\mathcal{C}$ are identified by thresholding:

$$\mathcal{H}_{\mathcal{C}}^{\ell} = \left\{ h \,\middle|\, A_{\mathcal{C}}^{\ell,h} > \lambda_{\mathcal{C}} \right\} \tag{11}$$

This unified formulation allows HAI to selectively suppress over-attending heads based on their attention distribution patterns.

Suppressing Over-Attention in System and Text Heads Based on Figure 3, we analyze each attention head in the LLaVA-1.5 decoder across layers for their focus on visual, textual, and system tokens. As shown in Figure 6c, visual attention heads are sparse, while system heads are highly redundant. Figure 7b further reveals text heads that overly attend to the previous token. These patterns indicate that certain heads consistently over-focus on language priors.

To address this issue, we introduce an attention suppression mechanism that downscales attention directed to prompt-like tokens. For each decoder layer ℓ and head h, we adjust the attention matrix $\mathbf{A}^{\ell,h}$ as follows:

$$\mathbf{A}_{i,j}^{\ell,h} = \begin{cases} (1 - \alpha_{\text{txt}}) \cdot \mathbf{A}_{i,j}^{\ell,h}, & \text{if } j \in \mathcal{I}_{\text{txt}},\ h \in \mathcal{H}_{\text{txt}} \\ (1 - \alpha_{\text{sys}}) \cdot \mathbf{A}_{i,j}^{\ell,h}, & \text{if } j \in \mathcal{I}_{\text{sys}},\ h \in \mathcal{H}_{\text{sys}} \end{cases} \tag{12}$$

To maintain a valid probability distribution, the adjusted weights are re-normalized:

$$\mathbf{A}_{i,j}^{\ell,h} = \frac{\mathbf{A}_{i,j}^{\ell,h}}{\sum_{j} \mathbf{A}_{i,j}^{\ell,h}}. \tag{13}$$

where α_{txt} and α_{sys} are suppression coefficients controlling the degree of attention reduction to prior text and system prompt tokens, respectively.

Experiments

Implementation and Experimental Setup

We implement VisFlow with a greedy decoding strategy (beam size = 1) and conduct all experiments on 8 NVIDIA RTX 4090 GPUs, with a maximum generation length of 64 tokens. Hyperparameters are configured as $\lambda_{\text{vis}} = 1$, $\lambda_{\text{sys}} = 0.8$, $\lambda_{\text{txt}} = 0.3$, and $\tau = \frac{1}{20}, \frac{1}{2}$ for salient and sink tokens, respectively. For LLaVA, we set $\alpha_{\text{sys}} = 0.6$ and $\alpha_{\text{txt}} = 1.0$; for MiniGPT4 and mPLUG-Owl2, which have shorter visual token sequences, we set $\alpha_{\text{sys}} = 0.4$ and

Methods	LLaVA-1.5			MiniGPT-4			mPLUG-Owl2		
	Random ↑	*Popular* ↑	*Adversarial* ↑	*Random* ↑	*Popular* ↑	*Adversarial* ↑	*Random* ↑	*Popular* ↑	*Adversarial* ↑
Greedy	81.54	76.53	73.54	77.56	67.50	69.11	83.90	77.30	74.82
Beam Search	82.64	79.34	78.15	78.54	70.20	71.62	87.33	81.42	78.95
OPERA [CVPR2024]	79.50	76.63	75.88	78.35	69.65	71.42	87.03	80.29	77.92
VCD [CVPR2024]	82.51	79.33	78.17	78.61	69.95	71.62	87.36	81.42	78.95
DoLa [ICLR2024]	82.81	79.47	78.36	80.23	73.00	73.23	87.90	81.53	79.18
PAI* [ECCV2024]	85.94	81.12	77.75	78.01	70.26	72.46	88.18	81.94	77.83
VAR [ICLR2025]	81.96	77.40	73.59	-	-	-	-	-	-
VisFlow (ours)	**89.55**	**87.09**	**84.35**	**80.86**	**73.61**	**73.94**	**88.72**	**82.19**	**80.17**

Table 1: Comparison of F1 scores on the POPE benchmark under three evaluation settings: Random, Popular, and Adversarial. PAI^* refers to the variant using exclusively the visual attention enhancement mechanism. Bold indicates the best results and higher F1-score indicate better. Results are averaged over five random runs.

Methods	LLaVA-1.5			MiniGPT-4			mPLUG-Owl2		
	$CHAIR_s$ ↓	$CHAIR_i$ ↓	Recall ↑	$CHAIR_s$ ↓	$CHAIR_i$ ↓	Recall ↑	$CHAIR_s$ ↓	$CHAIR_i$ ↓	Recall ↑
Greedy	20.0	6.8	59.1	25.0	9.2	**58.7**	23.0	9.6	**54.4**
Beam Search	20.0	6.9	57.0	24.0	9.2	56.7	18.0	6.4	53.0
OPERA [CVPR2024]	17.0	6.3	56.7	20.0	8.2	58.1	16.0	5.8	54.0
VCD [CVPR2024]	20.0	6.9	57.0	23.0	8.9	56.4	18.0	6.4	53.0
DoLa [ICLR2024]	19.0	6.5	57.0	19.0	8.1	56.3	18.0	6.1	53.0
VisFlow (ours)	**15.0**	**3.8**	**63.1**	**18.0**	**7.8**	57.3	**16.0**	**4.9**	53.0

Table 2: Comparison of CHAIR (instance-level $CHAIR_i$ and sentence-level $CHAIR_s$) and Recall scores on the MSCOCO dataset. Smaller $CHAIR_i$ and $CHAIR_s$ indicate less hallucinations. Results are averaged over five random runs.

$\alpha_{\text{txt}} = 0.6$. The parameters for POPE are $k = 20$ and $\delta = 20$, while for CHAIR, which requires fine-grained visual perception, we use $k = 10$ and $\delta = 0.4$. All baseline models are evaluated with their default configurations.

To improve efficiency, attention head types are identified once at the prefill stage and reused during decoding. TAI is applied from layer 2 onward, as early layers focus on global visual integration. HAI is applied across all layers for system heads, and to layers 0–7 for text heads to retain visual alignment and avoid over-reliance on language priors. For models with semantically compressed visual tokens (e.g., MiniGPT-4 (Chen et al. 2023a), mPLUG-Owl2 (Ye et al. 2024)), TAI is omitted due to limited effectiveness.

We conducted comprehensive evaluations of VisFlow on three widely used benchmarks: CHAIR (Rohrbach et al. 2018), POPE (Li et al. 2023b). These benchmarks collectively assess visual factuality, grounding robustness. Experimental results demonstrate that VisFlow consistently improves LVLM performance across diverse tasks, highlighting its effectiveness and generalizability. We compare VisFlow against several representative decoding-based methods, including VCD (Leng et al. 2024), DoLa (Chuang et al. 2023), OPERA (Huang et al. 2024), PAI (Liu, Zheng, and Chen 2024), and VAR (Kang et al. 2025).

Main Results

For the POPE benchmark, as shown in Table 1, we achieve the highest F1 scores across all subsets. The improvements are especially notable under the Adversarial setting, highlighting VisFlow's robustness against spurious correlations. As shown in Table 2, on the CHAIR, VisFlow yields substantially lower $CHAIR_s$ and $CHAIR_i$ scores compared to decoding-based baselines, indicating stronger alignment between generated captions and visual content.

More Analysis and Ablation Experiments

Component Ablation Study To assess the contribution of each component in VisFlow, we conduct ablation experiments on the CHAIR benchmark. As shown in Table 3, removing either TAI or HAI results in a noticeable drop in CHAIR scores. While removing HAI for system heads yields the lowest $CHAIR_s$, it reduces Recall, indicating a trade-off. In contrast, our full VisFlow achieves the best overall performance, with the lowest $CHAIR_i$ and highest Recall, confirming the effectiveness of the complete design.

Settings	**$CHAIR_s$** ↓	**$CHAIR_i$** ↓	**Recall** ↑
Greedy (Baseline)	20.0	6.8	59.1
w/o TAI	16.0	4.8	56.7
w/o HAI	16.0	5.3	58.4
w/o HAI for Txt. Heads	18.0	6.4	61.1
w/o HAI for Sys. Heads	**12.0**	4.0	60.1
Our Full VisFlow	15.0	**3.8**	**63.1**

Table 3: Ablation study on the CHAIR benchmark evaluating different components of VisFlow.

Sensitivity to Hyperparameters To determine the optimal enhancement strength, we vary the scaling factor k for salient visual tokens. As shown in Figure 8, the F1 Score peaks at $k = 20$, while both smaller ($k = 1$) and larger ($k = 30$) values degrade performance, indicating the importance

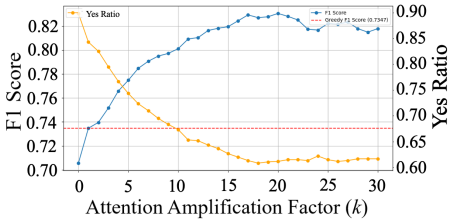


Figure 8: Effect of different k for salient tokens.

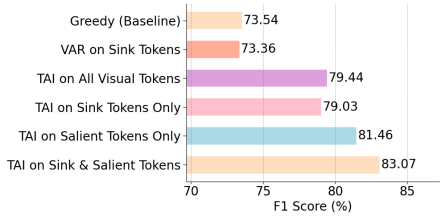


Figure 9: Comparison of token selection strategies.

of balanced visual emphasis. Figure 9 further shows that our selective enhancement strategy surpasses VAR (Kang et al. 2025), verifying that not all tokens contribute equally. We then tune the salient-token threshold τ in the TAI module on POPE (adv.) and CHAIR. As shown in Table 4, VisFlow consistently outperforms the baseline ($\tau = 1$), with $\tau = \frac{1}{20}$ achieving the best balance between hallucination suppression and generation quality.

Functional Analysis of Attention Heads To validate the functional roles of different attention head types, we identified the top-4 heads per type. We then performed causal interventions by zeroing them out from layer 0 and evaluated model performance on the CHAIR benchmark. As shown in Table 5, masking visual-sensitive heads notably increases CHAIR scores, confirming their critical role in visual alignment. Masking text-dominant heads in shallow layer mitigates hallucinations, whereas masking system-dominant heads shows minimal impact. These findings support the claim that not all heads are equally important.

Table 5 and Figures 10a–10b further show POPE-based interventions with thresholds λ_{sys}=0.8 and λ_{txt}=0.3. Masking text-dominant heads effectively reduces hallucinations, while masking system heads has negligible or even positive effects, suggesting redundancy.

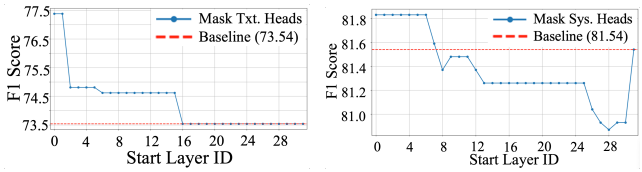
Decoding efficiency analysis We evaluate the inference efficiency of our method by comparing it with several baseline approaches in terms of Tokens Per Second (TPS) during decoding on the CHAIR Benchmark, as shown in Figure 11. Our approach achieves latency comparable to beam search, while being significantly faster than other methods, demonstrating superior efficiency.

Conclusion

We present VisFlow, a training-free and inference-time framework for VH in LVLMs. Our framework builds on the critical insight that not all tokens and heads are equally important for hallucination mitigation. Based on this, VisFlow directly intervenes in the attention dynamics within the decoder to improve visual alignment and reduce reliance on linguistic priors. Specifically, we propose a dual-level attention intervention approach: TAI: Enhances attention to critical visual tokens. HAI: Suppresses over-attention to non-visual tokens. Extensive experiments demonstrate that VisFlow significantly improves the visual faithfulness of generated responses. Our work provides new insights into decoding-time attention modulation as an effective means of reducing hallucination in LVLMs.

Settings (τ)	POPE (F1 ↑)	$\text{CHAIR}_{\text{BLEU-1}}$ ↑	CHAIR_{s} ↓
1 (Baseline)	73.54	28.9	20.0
1/10	82.81	29.1	15.0
1/20	**83.07**	**29.7**	**14.0**
1/30	82.14	29.1	14.0
0	79.44	29.3	15.0

Table 4: Sensitivity of salient-token threshold τ.



(a) Masking text heads. (b) Masking system heads.

Figure 10: Ablation study of HAI on POPE. (a) Evaluates the effect of masking text heads on POPE (adversarial); (b) The impact of masking system heads on POPE (random).

Setting	CHAIR_{s} ↓	CHAIR_{i} ↓	Recall ↑
Greedy (Baseline)	20.0	6.8	**59.1**
Mask Visual Heads	46.0	22.1	56.4
Mask System Heads	20.0	6.7	58.0
Mask Text Heads	19.0	7.2	58.4
Mask Text Heads (Shallow layer)	**15.0**	**5.7**	57.4
Mask Random Heads	20.0	7.4	56.0

Table 5: Comparison of masking different types of heads on the CHAIR benchmark.

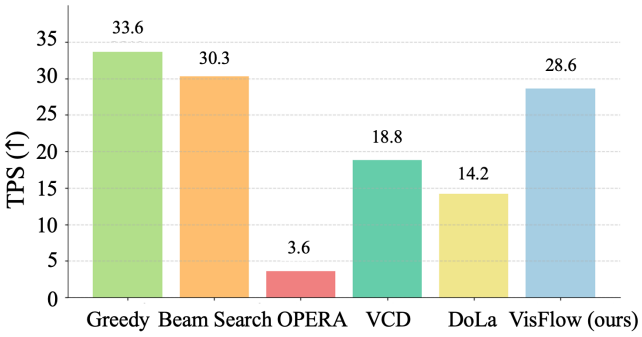


Figure 11: Comparison of our method with different baselines in terms of Tokens Per Second (TPS) during decoding on CHAIR Benchmark using LLaVA-1.5.

Acknowledgments

This work is supported by Guangdong Provincial Key Laboratory of Ultra High Definition Immersive Media Technology (Grant No. 2024B1212010006).

References

An, W.; Tian, F.; Leng, S.; Nie, J.; Lin, H.; Wang, Q.; Chen, P.; Zhang, X.; and Lu, S. 2025. Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 29915–29926.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1(2): 3.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Chen, C.; Liu, M.; Jing, C.; Zhou, Y.; Rao, F.; Chen, H.; Zhang, B.; and Shen, C. 2025. PerturboLLaVA: Reducing multimodal hallucinations with perturbative visual training. *arXiv preprint arXiv:2503.06486*.

Chen, J.; Zhu, D.; Shen, X.; Li, X.; Liu, Z.; Zhang, P.; Krishnamoorthi, R.; Chandra, V.; Xiong, Y.; and Elhoseiny, M. 2023a. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Chen, K.; Zhang, Z.; Zeng, W.; Zhang, R.; Zhu, F.; and Zhao, R. 2023b. Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic. *ArXiv*, abs/2306.15195.

Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Cui, E.; Zhu, J.; Ye, S.; Tian, H.; Liu, Z.; et al. 2024a. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 24185–24198.

Chen, Z.; Zhao, Z.; Luo, H.; Yao, H.; Li, B.; and Zhou, J. 2024c. Halc: Object hallucination reduction via adaptive focal-contrast decoding. *arXiv preprint arXiv:2403.00425*.

Chiang, W.-L.; Li, Z.; et al. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. https://vicuna.lmsys.org. Accessed: April 14, 2023.

Chuang, Y.-S.; Xie, Y.; Luo, H.; Kim, Y.; Glass, J.; and He, P. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*.

Darcet, T.; Oquab, M.; Mairal, J.; and Bojanowski, P. 2023. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*.

Fan, S.; Xie, L.; Shen, C.; Teng, G.; Yuan, X.; Zhang, X.; Huang, C.; Wang, W.; He, X.; and Ye, J. 2025. Improving complex reasoning with dynamic prompt corruption: A soft prompt optimization approach. *arXiv preprint arXiv:2503.13208*.

Gunjal, A.; Yin, J.; and Bas, E. 2024. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18135–18143.

Huang, Q.; Dong, X.; Zhang, P.; Wang, B.; He, C.; Wang, J.; Lin, D.; Zhang, W.; and Yu, N. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13418–13427.

Huo, F.; Xu, W.; Zhang, Z.; Wang, H.; Chen, Z.; and Zhao, P. 2024. Self-introspective decoding: Alleviating hallucinations for large vision-language models. *arXiv preprint arXiv:2408.02032*.

Jiang, C.; Xu, H.; Dong, M.; Chen, J.; Ye, W.; Yan, M.; Ye, Q.; Zhang, J.; Huang, F.; and Zhang, S. 2024. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27036–27046.

Kang, S.; Kim, J.; Kim, J.; and Hwang, S. J. 2025. See what you are told: Visual attention sink in large multimodal models. *arXiv preprint arXiv:2503.03321*.

Leng, S.; Zhang, H.; Chen, G.; Li, X.; Lu, S.; Miao, C.; and Bing, L. 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*, 13872–13882.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W. X.; and Wen, J.-R. 2023b. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, H.; Xue, W.; Chen, Y.; Chen, D.; Zhao, X.; Wang, K.; Hou, L.; Li, R.; and Peng, W. 2024b. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

Liu, S.; Ye, H.; Xing, L.; and Zou, J. 2024c. Reducing hallucinations in vision-language models via latent space steering. *arXiv preprint arXiv:2410.15778*.

Liu, S.; Zheng, K.; and Chen, W. 2024. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, 125–140. Springer.

Manakul, P.; Liusie, A.; and Gales, M. J. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

Michel, P.; Levy, O.; and Neubig, G. 2019. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32.

Park, Y.; Lee, D.; Choe, J.; and Chang, B. 2025. Convis: Contrastive decoding with hallucination visualization for mitigating hallucinations in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 6434–6442.

Rohrbach, A.; Hendricks, L. A.; Burns, K.; Darrell, T.; and Saenko, K. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.

Ru, J.; Xie, Y.; Zhuang, X.; Yin, Y.; and Zou, Y. 2025. Do we really have to filter out random noise in pre-training data for language models? *arXiv preprint arXiv:2502.06604*.

Sarkar, P.; Ebrahimi, S.; Etemad, A.; Beirami, A.; Arık, S. Ö.; and Pfister, T. 2024. Data-augmented phrase-level alignment for mitigating object hallucination. *arXiv preprint arXiv:2405.18654*.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Su, J.; Ahmed, M.; Lu, Y.; Pan, S.; Bo, W.; and Liu, Y. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568: 127063.

Sun, Z.; Zang, X.; Zheng, K.; Song, Y.; Xu, J.; Zhang, X.; Yu, W.; and Li, H. 2024. Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. *arXiv preprint arXiv:2410.11414*.

Tang, F.; Huang, Z.; Liu, C.; Sun, Q.; Yang, H.; and Lim, S.-N. 2025. Intervening anchor token: Decoding strategy in alleviating hallucinations for MLLMs. In *The Thirteenth International Conference on Learning Representations*.

Team, Q. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wang, L.; Li, L.; Dai, D.; Chen, D.; Zhou, H.; Meng, F.; Zhou, J.; and Sun, X. 2023. Label words are anchors: An information flow perspective for understanding in-context learning. *arXiv preprint arXiv:2305.14160*.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, X.; Pan, J.; Ding, L.; and Biemann, C. 2024b. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. *arXiv preprint arXiv:2403.18715*.

Xing, Y.; Li, Y.; Laptev, I.; and Lu, S. 2024. Mitigating object hallucination via concentric causal attention. *Advances in Neural Information Processing Systems*, 37: 92012–92035.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yang, T.; Li, Z.; Cao, J.; and Xu, C. 2025b. Understanding and Mitigating Hallucination in Large Vision-Language Models via Modular Attribution and Intervention. In *International Conference on Learning Representations (ICLR)*. Poster.

Yang, Z.; Luo, X.; Han, D.; Xu, Y.; and Li, D. 2025c. Mitigating hallucinations in large vision-language models via dpo: On-policy data hold the key. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 10610–10620.

Ye, Q.; Xu, H.; Ye, J.; Yan, M.; Hu, A.; Liu, H.; Qian, Q.; Zhang, J.; and Huang, F. 2024. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 13040–13051.

Yin, H.; Si, G.; and Wang, Z. 2025. ClearSight: Visual Signal Enhancement for Object Hallucination Mitigation in Multimodal Large Language Models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 14625–14634.

Yin, S.; Fu, C.; Zhao, S.; Xu, T.; Wang, H.; Sui, D.; Shen, Y.; Li, K.; Sun, X.; and Chen, E. 2024. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12): 220105.

Zhang, Z.; Yadav, S.; Han, F.; and Shutova, E. 2024. Cross-modal Information Flow in Multimodal Large Language Models. *arXiv preprint arXiv:2411.18620*.

Zhuang, X.; Xie, Y.; Deng, Y.; Liang, L.; Ru, J.; Yin, Y.; and Zou, Y. 2025a. VARGPT: Unified Understanding and Generation in a Visual Autoregressive Multimodal Large Language Model. *arXiv preprint arXiv:2501.12327*.

Zhuang, X.; Zhu, Z.; Xie, Y.; Liang, L.; and Zou, Y. 2025b. VASparse: Towards Efficient Visual Hallucination Mitigation for Large Vision-Language Model via Visual-Aware Sparsification. *arXiv preprint arXiv:2501.06553*.

Zou, X.; Wang, Y.; Yan, Y.; Lyu, Y.; Zheng, K.; Huang, S.; Chen, J.; Jiang, P.; Liu, J.; Tang, C.; et al. 2024. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. *arXiv preprint arXiv:2410.03577*.