Taming the Phantom: Token-Asymmetric Filtering for Hallucination Mitigation in Large Vision-Language Models

Shuyi Ouyang[1,2], Hongyi Wang[1], Gongfan Fang[2], Xinyin Ma[2], Lanfen Lin[1*], Xinchao Wang[2*]

[1]Zhejiang University
[2]National University of Singapore
{oysy, llf}@zju.edu.cn, xinchao@nus.edu.sg

Abstract

Hallucination in Large Vision-Language Models (LVLMs) remains a critical challenge, undermining their reliability in real-world applications. Existing studies have investigated the causes of hallucination at the modality level and proposed effective strategies. However, interaction patterns beyond the modality level remain insufficiently explored. In this paper, we conduct a token-level analysis and identify two key phenomena: (1) a small subset of textual tokens in LVLMs exert disproportionate influence in the visual-active layers, surpassing that of the visual modality and potentially misleading visual understanding; (2) while LVLMs can correctly identify key visual information, insufficient focus on these cues can sometimes lead to hallucinations. Based on such observation, we attribute hallucinations in LVLMs to two token-level causes: the disproportionate influence of certain textual tokens (phantom tokens) and the underutilization of critical visual cues (anchor tokens). To mitigate these issues, we introduce Token-Asymmetric Filtering (TAF)—a training-free, plug-and-play method that modulates intermediate attention maps in LVLMs. TAF isolates the influence of phantom tokens and emphasizes the influence of anchor tokens in the visual-active layers. Experimental results across multiple benchmarks demonstrate that TAF significantly mitigates hallucinations across a range of state-of-the-art LVLMs.

1 Introduction

Large Vision-Language Models (LVLMs) have demonstrated remarkable capabilities in a variety of multi-modal tasks, including visual question answering, image captioning, and open-ended visual dialogue (Chen et al. 2023b; Li et al. 2023a; Liu et al. 2023b; Zhou et al. 2022; Zhu et al. 2023). However, despite these impressive advances, hallucination—i.e., the generation of outputs that are unfaithful to the visual input—remains a persistent and critical challenge. Such a phenomenon can severely undermine the trustworthiness of LVLMs in safety-critical or knowledge-intensive applications (Hu et al. 2023; Wang et al. 2023; Chen et al. 2024b; Liu et al. 2023d).

To mitigate the hallucination problem in LVLMs (Agrawal, Batra, and Parikh 2016; Agarwal, Shetty, and Fritz 2020; Biten, Gómez, and Karatzas 2022), supervised fine-tuning approaches have garnered significant attention, often relying on large-scale annotated datasets and incurring substantial training costs (Zhao et al. 2023; Yu et al. 2024; Sun et al. 2023; Liu et al. 2023a; Lyu et al. 2024). As more efficient alternatives, inference-time strategies such as contrastive decoding and attention-based modulation have been proposed (An et al. 2024; Xing et al. 2024; Huo et al. 2024; Gong et al. 2024; Zhao et al. 2023).

Recent studies have approached hallucination at the modality level, examining how imbalance or misalignment between the visual and linguistic modalities leads to factual inconsistency in LVLMs (Chen et al. 2024a; Yin, Si, and Wang 2025; Jiang et al. 2024). While prior studies have advanced understanding of hallucinations in vision-language models, they overlook fine-grained token-level behaviors, limiting the flexibility and effectiveness of hallucination mitigation. We address this gap by conducting a token-level analysis that identifies more precise and actionable origins of hallucinations in LVLMs.

We employed the saliency analysis to observe the behavior of individual tokens in the layers of LVLMs, abstracting the trends as shown in Figure 1(a). Our observations reveal that, in hallucination scenarios, certain textual tokens exert disproportionately misleading influence on the model's visual understanding, dominating the attention dynamics by several times more than the visual tokens. These tokens pose a risk of interfering with the visual context in the intermediate layers, distorting the model's perception of the image. We term these disruptive elements *phantom tokens*. Meanwhile, as shown in Figure 1(b), we observe that a few *anchor tokens* from the visual modality are sufficient to steer the model's attention toward semantically meaningful regions of the image. However, when visual attention is dispersed or misaligned, the model may fail to extract the necessary evidence, which can result in hallucination. These observations reveal a token-level asymmetry: (1) while the language modality contributes interpretative flexibility for reasoning, a small number of phantom tokens can disproportionately influence and misguide the model's visual understanding; and (2) the visual modality provides factual grounding, only a small number of anchor tokens often carry the essential visual evidence required for accurate grounding.

Based on the above analysis, we identify two token-level causes of hallucination in LVLMs: (1) misleading reasoning

*Corresponding authors.

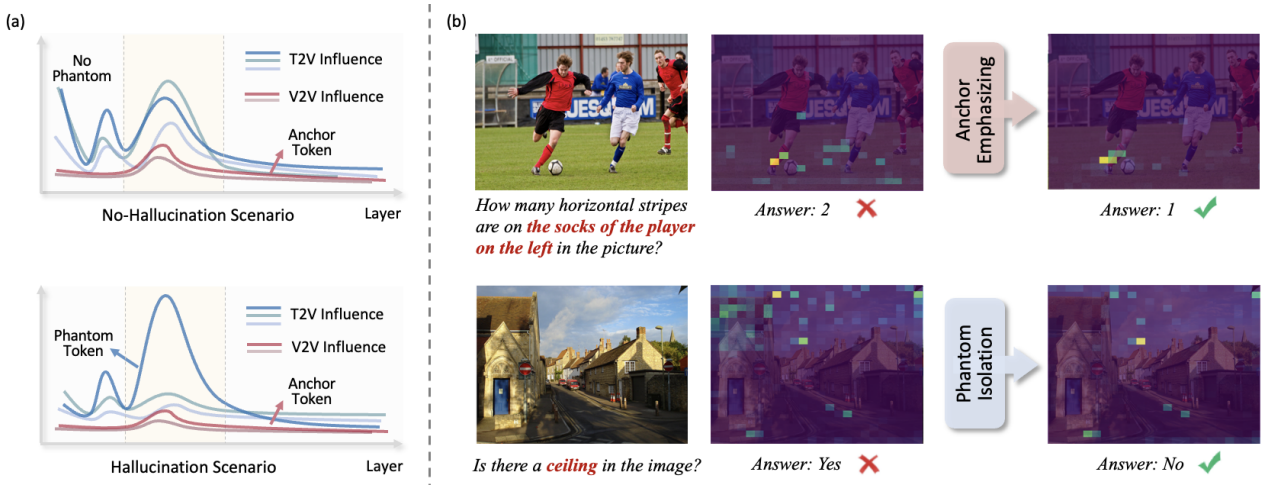

Figure 1: (a) Token-level interaction analysis. Each curve corresponds to the influence of one token. Text-to-Vision (T2V) curves and Vision-to-Vision (V2V) curves represent the influence of textual and visual tokens, respectively, on visual understanding. Visual understanding primarily emerges in the intermediate layers of the LVLM, highlighted in yellow. (b) Example results and corresponding attention map visualizations. In the first row, emphasizing anchor tokens enables the model to focus more effectively on critical regions of the image while suppressing attention to irrelevant areas, thereby facilitating a clearer and more grounded visual understanding. In the second row, isolating phantom tokens prevents the model from mistakenly interpreting the "roof" as a "ceiling," effectively suppressing unreliable textual influence on the visual understanding.

induced by phantom tokens in the language modality, and (2) insufficient focus on key visual evidence due to underutilization of anchor tokens. To mitigate these issues, we propose Token-Asymmetric Filtering (TAF)—a plug-and-play, training-free mechanism that dynamically modulates attention maps in the visual-active layers of LVLMs. TAF identifies phantom and anchor tokens based on the influence between modalities. As shown in Figure 2(b), TAF mitigates hallucinations from two perspectives: it isolates the influence of phantom tokens on visual attention while emphasizing anchor tokens to guide the model's focus toward relevant visual content.

By rebalancing token contributions without retraining, TAF mitigates hallucination and improves interpretability. Experiments on multiple benchmarks and LVLM architectures show that TAF consistently improves factual alignment, confirming its effectiveness and generality.

Our main contributions are summarized as follows:

1. We propose a token-level analysis of hallucinations in LVLMs, utilizing saliency scores to analyze the interactions between visual and textual tokens in the intermediate layers of the model.
2. We attribute hallucinations in LVLMs to two token-level causes: the disproportionate influence of certain textual tokens and the underutilization of critical visual tokens.
3. We propose Token-Asymmetric Filtering (TAF), a training-free method that modulates attention maps by isolating the influence of phantom tokens and emphasizing the influence of anchor tokens.
4. Extensive experiments across multiple LVLM architectures and benchmarks demonstrate the effectiveness and generality of TAF in mitigating hallucination.

2 Related Works

2.1 Large Vision-Language Models

Unlike earlier models such as BLIP (Li et al. 2022, 2023a) and BERT-based vision-language models (Devlin et al. 2019; Liu et al. 2019), Large Vision-Language Models (LVLMs) have significantly enhanced their understanding capabilities by integrating the rapidly advancing Large Language Models (LLMs), enabling them to perform complex vision-language reasoning tasks (Bai et al. 2023; Grattafiori et al. 2024; Liu et al. 2024; Chen et al. 2024c, 2023a). In general, recent LVLMs consist of three components: a visual encoder, a connector, and a pre-trained LLM. Researchers connect the visual encoder with the LLM through various connector modules, such as linear projections (Liu et al. 2023b) or Q-formers (Zhu et al. 2023). These models typically undergo two training phases: a pre-training phase and a fine-tuning phase. Early attempts, such as Flamingo (Alayrac et al. 2022), Gemini (Team et al. 2023), and BLIP-2 (Li et al. 2023a), have already demonstrated promising results. More recent studies, such as LLaVA-v1.5 (Liu et al. 2024), Qwen2.5-VL (Wang et al. 2024), and xGen-MM (Xue et al. 2024), have further advanced the field and significantly enhanced model capabilities. To further improve vision-language representation abilities, the latest research includes using higher-resolution visual encoders, larger and more powerful LLMs, and methods such as Reinforcement Learning with Human Feedback (RLHF) (Ouyang et al. 2022; Yu et al. 2024).

2.2 Hallucination Mitigation in LVLMs

In the field of Natural Language Processing (NLP), hallucinations are typically defined as the phenomenon of generat-

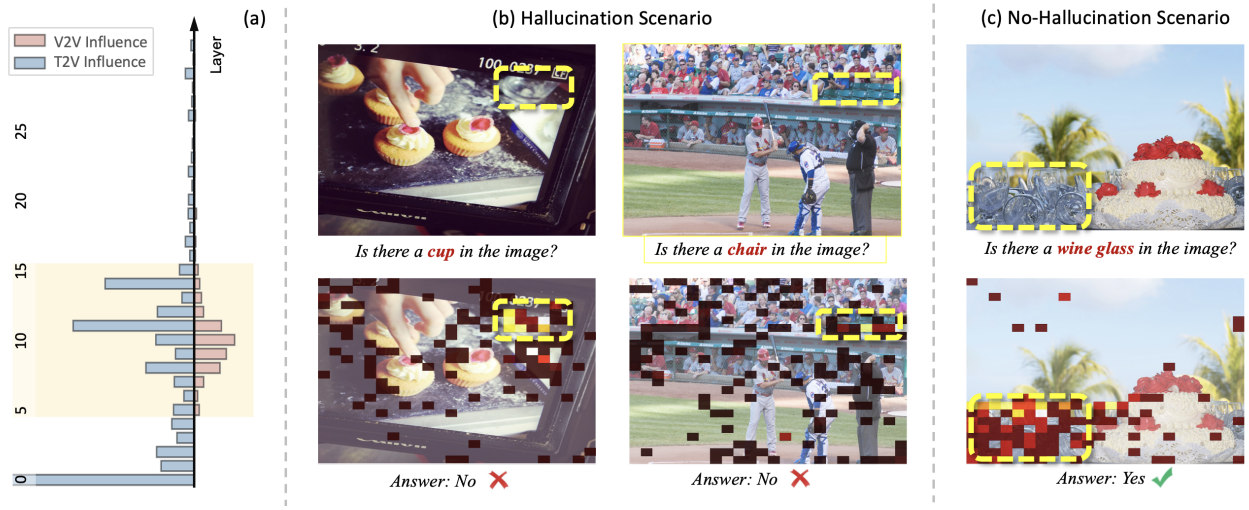

Figure 2: (a) Layer-wise visualization of token influence across the Transformer layers of LLaVA-1.5-7B. (b) and (c) show the saliency maps corresponding to hallucination and no-hallucination scenarios, respectively.

ing erroneous or meaningless content (Ji et al. 2023; Weijia et al. 2023). In the context of LVLMs, object hallucination refers to the occurrence where the text generated by the model is semantically coherent, but inconsistent with the actual objects present in the accompanying image (Huang et al. 2024b; Rohrbach et al. 2018; Wu et al. 2024). Existing studies have identified several potential causes of hallucinations, including data biases and misalignment between visual and linguistic information, and have proposed corresponding strategies. One approach involves fine-tuning the model using preference-based data, further calibrating the model's behavior by introducing high-quality labeled data (Li et al. 2024; Yu et al. 2024; Pi et al. 2024; Zhou et al. 2024), although this strategy requires significant computational resources. Another strategy captures richer visual details using auxiliary inputs such as depth and segmentation maps (Jain, Yang, and Shi 2024; Lee et al. 2024; Zhao et al. 2023), though these approaches lack cross-task generalizability. Other methods introduce additional visual models (Biten, Gómez, and Karatzas 2022; Yin et al. 2024). Recent research improves the decoding process by inducing hallucinations through transformations like blurring, rotation, and cropping of the original visual input, and penalizing hallucination tokens during decoding (Chen et al. 2024d; Chuang et al. 2023; Leng et al. 2024; Woo et al. 2024; Zhong et al. 2024), though this increases inference time. Recently, some studies (An et al. 2024; Liu, Zheng, and Chen 2024; Jiang et al. 2024; Yin, Si, and Wang 2025) have explored attention mechanisms. Existing research remained at the modality level. We explore token-level behaviors during inference and modulate them to effectively mitigate hallucination.

3 Token-Level Origins of Hallucinations

3.1 Measuring Token-Level Influence

We employ the representative saliency-based technique to measure the influence of individual tokens across different layers of the model. We define $I^{l,j}_{\text{v2v}}$ to denote the Vision-to-Vision (V2V) influence of the j-th visual token at l-th layer, while $I^{l,k}_{\text{t2v}}$ represents the Text-to-Vision (T2V) influence of the k-th textual token on visual tokens at the same layer. The computation of both quantities is given by the following:

$$I^{l,j}_{\text{v2v}} = \frac{1}{N_{\mathcal{V}}} \sum_i \sum_h A^l_{h,i,j} \odot \frac{\partial \mathcal{L}}{\partial A^l_{h,i,j}}, \quad i \in \mathcal{V}, j \in \mathcal{V}, \quad (1)$$

$$I^{l,k}_{\text{t2v}} = \frac{1}{N_{\mathcal{V}}} \sum_i \sum_h A^l_{h,i,k} \odot \frac{\partial \mathcal{L}}{\partial A^l_{h,i,k}}, \quad i \in \mathcal{V}, k \in \mathcal{T}, \quad (2)$$

where the set $\mathcal{V}$ refers to the indices of visual tokens, and $\mathcal{T}$ denotes the indices of textual tokens. $N_{\mathcal{V}}$ is the number of visual tokens and serves as a normalization factor. Let $A^l_{h,i,j}$ denote the normalized attention weight from the i-th query token to the j-th key token in the h-th head of layer l. The gradient term $\partial \mathcal{L} / \partial A^l_{h,i,j}$ measures the sensitivity of the loss $\mathcal{L}$ to each attention weight. The operator $\odot$ denotes the element-wise product, and the summation aggregates the results into a single scalar value.

To analyze the distinct behaviors of the two modalities in the intermediate layers of the model, we first examine the average token-level saliency for each modality at every layer. Specifically, we define I^l_{v2v} and I^l_{t2v} as the mean scores for the V2V influence and T2V influence at layer l, respectively. These values are computed as follows:

$$I^l_{\text{v2v}}, I^l_{\text{t2v}} = \frac{1}{N_{\mathcal{V}}} \sum_j I^{l,j}_{\text{v2v}}, \frac{1}{N_{\mathcal{T}}} \sum_k I^{l,k}_{\text{t2v}}, \quad j \in \mathcal{V}, k \in \mathcal{T}, \quad (3)$$

where $N_{\mathcal{T}}$ is the number of textual tokens. Using LLaVA-1.5-7B as an example, Figure 2(a) visualizes the average token-level influence across layers. It reveals a clear asymmetry between the two modalities: the T2V influence dominates most layers, while the V2V influence is largely confined to a limited range of intermediate layers. Layers with active V2V influence are defined as *visual-active layers*.

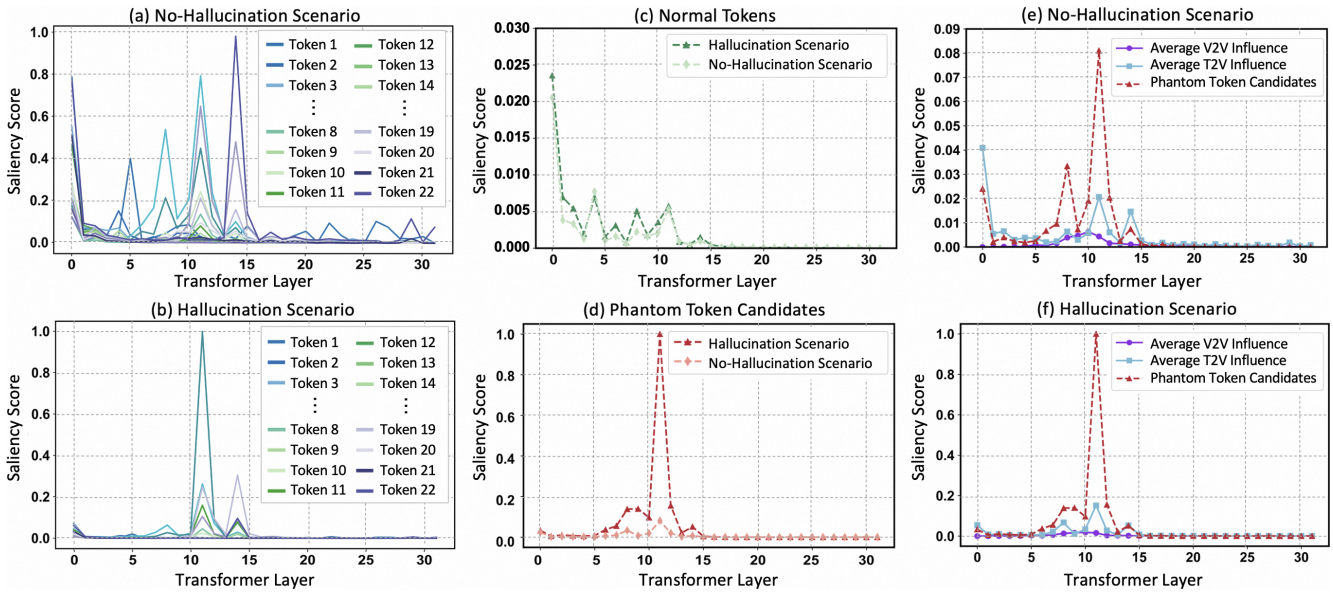

Figure 3: Manifold curvature visualization of T2V and V2V Influence. (a) and (b) compare the T2V Influence of each textual token under no-hallucination and hallucination scenarios. (c) and (d) compare normal tokens with phantom candidates under both conditions. (e) and (f) contrast the T2V Influence of phantom token candidates with the average T2V and V2V Influence.

In LVLMs, the visual modality is primarily responsible for providing factual grounding, while the language modality contributes to logical reasoning. An excessive dominance of the language modality risks overwhelming or misguiding the visual signal, potentially leading to hallucinations. Therefore, it is crucial to leverage the visual modality efficiently within the visual-active layers to preserve factual consistency and enhance multimodal alignment.

3.2 Visual Attention Dispersion

In Figures 2(b) and 2(c), we visualize the saliency maps of visual tokens in the intermediate layers. Leveraging the powerful understanding capability of the LLM, the model often identifies the key regions in the image most relevant to the answer (with the highest saliency values shown as yellow blocks). Comparing the hallucination scenario in (b) with the no-hallucination scenario in (c), saliency in (c) is more focused on the critical regions. Therefore, we infer that while LVLMs know where to look, a lack of sufficient focus on these regions may be a potential cause of hallucinations.

3.3 Textual Token Asymmetry

Textual tokens, being central to generation, attract stronger query attention than visual tokens. To better understand hallucination-related dynamics, we analyze the saliency of individual textual tokens across layers, specifically focusing on the T2V influence $I_{\text{t2v}}^{l,k}$. The visualization results are shown in Figure 3. As shown in Figure 3(a) and (b), compared to no-hallucination cases, the hallucination examples exhibit the emergence of certain tokens with abnormally high influence. The saliency values of these tokens in the visual-active layers exceed those of visual tokens by several folds, suggesting a strong potential to mislead the visual context. We refer to such tokens as *phantom tokens*.

As illustrated in Figure 3(c) and (d), to further investigate this phenomenon, we compare the phantom token candidates with normal tokens, whose influence levels are close to the average. In Figure 3(e) and (f), we also include the average V2V and T2V influence scores as references. Phantom tokens show overwhelmingly dominant influence across multiple layers. To prevent phantom tokens from overshadowing or distorting visual facts, suppressing or decoupling them in visual-active layers is necessary.

3.4 Insights

The preceding analysis, using LLaVA-1.5-7B as an example, reveals key dynamics in LVLMs. The activation range of the visual modality is narrow, primarily providing factual information within a limited set of intermediate layers. Furthermore, there are two potential token-level causes of hallucinations in LVLMs: (1) the model knows where to look but lacks sufficient focus, leading to inadequate attention on critical visual tokens; (2) a small subset of textual tokens exhibits disproportionately high influence in the visual-active layers, potentially introducing misleading linguistic noise. Similar phenomena have been observed in other LVLM backbones, and further analyses of additional LVLMs are provided in the Appendix.

These observations underscore the need for a targeted mechanism that both promotes the effective utilization of visual evidence and mitigates the disruptive effect of over-dominant textual tokens.

4 Method

Based on the above analysis, we propose a plug-and-play, training-free hallucination mitigation strategy, Token-

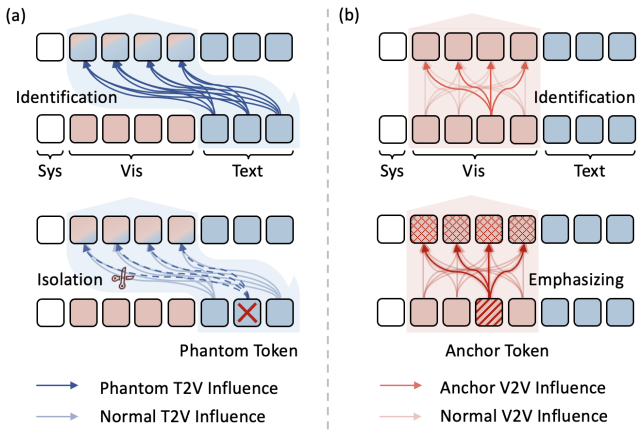

Figure 4: (a) and (b) illustrate the identification and filtering of phantom and anchor tokens, respectively.

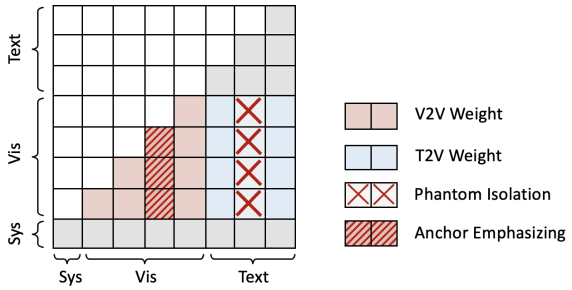

Figure 5: Illustration of the proposed TAF mechanism applied to the attention map.

Asymmetric Filtering (TAF), as illustrated in Figure 4, 5. TAF aims to suppress disruptive T2V signals and enhance the critical visual cues within the visual-active layers.

4.1 Identifying Phantom and Anchor Tokens

We first identify two types of critical tokens: (1) *phantom tokens* (set $\mathcal{P}$): textual tokens that exert abnormally high influence on visual context; (2) *anchor tokens* (set $\mathcal{A}$): visual tokens that serve as important reference points for factual alignment.

To compute the influence signals used for token identification, we define the average V2V and T2V attention scores at each layer l as:

$$V^{l,j} = \frac{1}{N_\mathcal{V}} \sum_i A^l_{i,j}, \quad i \in \mathcal{V}, j \in \mathcal{V}, \tag{4}$$

$$T^{l,k} = \frac{1}{N_\mathcal{V}} \sum_i A^l_{i,k}, \quad i \in \mathcal{V}, k \in \mathcal{T}, \tag{5}$$

where $A^l_{i,j}$ denotes the attention weight from the i-th token to the j-th token, and $N_\mathcal{V}$ is the number of visual tokens.

Anchor Tokens. We define the anchor score, $S^{l,j}_\mathcal{A} = V^{l,j}$, as the average V2V attention scores received by the j-th visual token from all other visual tokens. Anchor tokens are classified with a dynamic criterion, allowing automatic adaptation to the score distribution:

$$\mathcal{A}^l = \{j \in \mathcal{V} \mid S^{l,j}_\mathcal{A} \geq \mu^l_\mathcal{A} + \lambda_\mathcal{A} \cdot \sigma^l_\mathcal{A}\}, \tag{6}$$

where $\lambda_\mathcal{A} = 1$ by default, $\mu^l_\mathcal{A}$ and $\sigma^l_\mathcal{A}$ denote the mean and standard deviation of $\{S^{l,j}_\mathcal{A}\}_{j \in \mathcal{V}}$ at layer l.

We define the maximum value of $S^{l,j}_\mathcal{A}$ at layer l as $\Lambda^l_\mathcal{A}$, and the maximum value across all layers as $\Lambda_\mathcal{A}$. The set of *visual-active layers* L_v is defined as those layers l satisfying $\mu^l_\mathcal{A} > 0.35\Lambda_\mathcal{A}$.

Phantom Tokens. We define the phantom score $S^{l,k}_\mathcal{P}$ by comparing a token's T2V attention scores against the average score $\mu^l_\mathcal{A}$:

$$S^{l,k}_\mathcal{P} = \frac{T^{l,k}}{\mu^l_\mathcal{A} + \epsilon}, \quad j \in \mathcal{V}, k \in \mathcal{T}, \tag{7}$$

where ϵ is a small constant to prevent division by zero. Unlike anchor tokens, phantom tokens should only be identified when the attention distribution exhibits extreme outliers. To achieve this, we employ a dual-condition thresholding strategy. For each layer l, $S^{l,k}_\mathcal{P}$ have a mean of $\mu^l_\mathcal{P}$ and a standard deviation of $\sigma^l_\mathcal{P}$. Only when $\sigma^l_\mathcal{P} \geq 1.5 \cdot \sigma^l_\mathcal{A}$ do we apply the adaptive thresholding rule:

$$\mathcal{P}^l = \{k \in \mathcal{T} \mid S^{l,k}_\mathcal{P} \geq \mu^l_\mathcal{P} + \lambda_\mathcal{P} \cdot \sigma^l_\mathcal{P}\}. \tag{8}$$

Otherwise, we set $\mathcal{P}^l = \emptyset$. We set $\lambda_\mathcal{P} = 1.5$ by default. This design ensures that phantom tokens are only assigned in genuinely extreme cases, consistent with their definition as rare but disproportionately influential tokens.

4.2 Modulating Token Influence via Asymmetric Filtering

Having identified the sets of anchor tokens $\mathcal{A}$ and phantom tokens $\mathcal{P}$, we introduce *Anchor Emphasizing* and *Phantom Isolation* to selectively filter the signals. Given our primary objective to maintain the integrity and focus on critical visual facts, we restrict this filtering to the visual-active layers L_v. During inference, we adjust the unnormalized attention logits ξ^l_h rather than the normalized A^l_h. Specifically, the attention score is formulated as the following:

$$A^l_h = \text{softmax}(\xi^l_h) = \text{softmax}\left(\frac{Q^l_h K^{l\top}_h}{\sqrt{d}}\right), \tag{9}$$

where ξ^l_h is computed as the scaled dot product between the query and key matrices: Q^l_h and K^l_h, respectively. The dimension d refers to the size of the query / key vectors.

Anchor Emphasizing strengthens attention to anchor tokens by positively adjusting the attention logits ξ^l_h along pathways from these anchors to all visual tokens. Phantom Isolation operates by decoupling attention logits ξ^l_h associated with the attention pathways from phantom tokens to all visual tokens. For each attention head h at layer $l \in L_v$, we apply the following adjustment:

$$\xi^l_h = \xi^l_h + \alpha \cdot M^l_{\mathcal{A},h} \circ \xi^l_h - M^l_{\mathcal{P},h} \circ \xi^l_h, \tag{10}$$

where $\circ$ denotes element-wise multiplication, and α is a positive scaling coefficient used to emphasize anchor tokens. The binary masks $M^l_{\mathcal{A},h}$ and $M^l_{\mathcal{P},h}$ specify attention paths

Methods	LLaVA-1.5			Qwen-VL			Qwen2.5-VL		
	Random	*Popular*	*Adversarial*	*Random*	*Popular*	*Adversarial*	*Random*	*Popular*	*Adversarial*
Regular	82.63	79.12	76.84	82.95	79.66	77.13	83.21	80.17	78.74
DoLa (Chuang et al. 2023)	84.13	80.89	76.33	84.78	81.25	77.82	83.89	80.45	79.57
OPERA (Huang et al. 2024a)	86.46	83.27	80.24	86.58	83.68	80.64	85.81	84.24	81.22
VCD (Leng et al. 2024)	86.45	83.01	78.55	86.94	83.59	80.03	86.41	84.56	82.33
AGLA (An et al. 2024)	87.13	83.92	81.42	87.69	84.28	83.14	88.45	86.13	84.96
ClearSight (Yin, Si, and Wang 2025)	88.69	84.68	81.04	88.53	85.01	81.43	89.24	85.79	82.57
Ours	**90.47**	**88.03**	**86.21**	**90.72**	**88.74**	**87.69**	**91.18**	**89.52**	**88.47**

Table 1: Comparison of the average F1-score evaluation results on the offline POPE benchmark. Higher F1-scores indicate better performance, with the best results highlighted in bold.

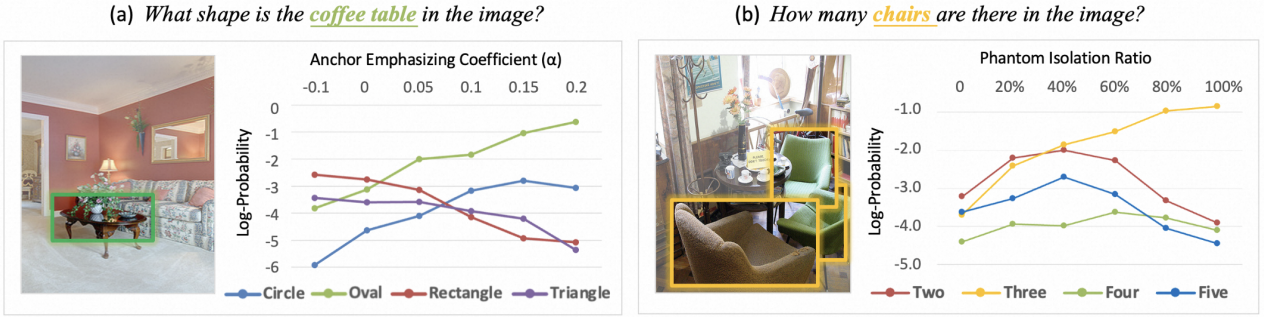

Figure 6: Effect of (a) anchor emphasizing and (b) phantom isolation. In (a), anchor emphasizing corrects the predicted shape of the table from *"rectangle"* to *"oval."* In (b), phantom isolation revises the predicted number of chairs from *"two"* to *"three."* The Phantom Isolation Ratio quantifies the proportion of attention logits suppressed for identified phantom tokens.

associated with anchor and phantom tokens, respectively. These are defined as:

$$M^l_{\mathcal{A},h}(i,j) = \mathbb{I}(i \in \mathcal{V}, j \in \mathcal{A}), \tag{11}$$

$$M^l_{\mathcal{P},h}(i,k) = \mathbb{I}(i \in \mathcal{V}, k \in \mathcal{P}), \tag{12}$$

where $\mathbb{I}(\cdot)$ is the indicator function. The attention enhancement of anchor tokens encourages the model to more effectively focus on critical visual cues, strengthening attention to semantically significant image regions. The influence of phantom tokens to the visual modality is cut off in the attention maps, thereby suppressing the introduction of textual noise in the visual-active layers without disrupting the factual cross-modal alignment.

5 Experiment

5.1 Benchmarks

Following the practice in prior works (Chen et al. 2024d; Leng et al. 2024), we evaluate the hallucination mitigation performance of our proposed Token-Asymmetric Filtering (TAF) on three widely adopted benchmarks: the offline Polling-based Object Probing Evaluation (POPE) (Li et al. 2023b; Chen et al. 2024d), Caption Hallucination Assessment with Image Relevance (CHAIR) (Rohrbach et al. 2018) on MSCOCO dataset (Lin et al. 2014), and general-purposed Multimodal Large Language Model Evaluation (MME) benchmark (Fu et al. 2024).

POPE formulates hallucination detection as a binary classification task, where models are prompted with simple yes/no questions such as "Is there a chair in the image?". Following prior research (Chen et al. 2024d), we primarily evaluate hallucination-related behaviors using the F1-score.

The CHAIR benchmark measures hallucination in image captioning from both sentence-level (CHAIR_s) and instance-level (CHAIR_i) perspectives. Recall is also reported to assess the model's ability to correctly identify and describe visual entities.

MME is a comprehensive benchmark that includes ten perception-related tasks and four cognition-related tasks, with all tasks evaluated using accuracy. The results of the MME benchmark are presented in the Appendix.

5.2 LVLM Backbones

To assess the generalizability and effectiveness of our TAF, we conduct experiments on LVLMs from three distinct cross-modal architectures: linear projection, Q-Former-based bridging, and single-tower Transformer. The models utilizing linear projection include LLaVA (Liu et al. 2023c) and the Qwen-VL (Bai et al. 2023; Wang et al. 2024) series, while MiniGPT-4 (Chen et al. 2023a) and InstructBLIP (Dai et al. 2023) use Q-Former, and mPLUG-Owl2 (Ye et al. 2024) employs a single-tower Transformer. Following prior works (Yin, Si, and Wang 2025; Leng et al. 2024), we evaluate our approach on a set of representative models, including LLaVA-1.5-7B (Liu et al. 2023c), Qwen-VL-7B (Bai et al.

Methods	LLaVA-1.5			MiniGPT-4			mPLUG-Owl2		
	$CHAIR_s \downarrow$	$CHAIR_i \downarrow$	Recall ↑	$CHAIR_s \downarrow$	$CHAIR_i \downarrow$	Recall ↑	$CHAIR_s \downarrow$	$CHAIR_i \downarrow$	Recall ↑
Greedy	47.6	15.1	76.6	39.5	15.6	55.8	53.7	17.9	72.3
OPERA (Huang et al. 2024a)	47.8	14.2	77.1	40.3	15.4	57.6	56.2	18.0	70.3
VCD (Leng et al. 2024)	51.4	15.8	76.0	39.9	16.2	54.2	64.1	19.2	71.6
DoLa (Chuang et al. 2023)	49.5	14.9	77.2	39.6	16.0	58.1	54.1	16.9	71.4
AGLA (An et al. 2024)	46.6	14.5	76.5	39.1	14.7	57.2	53.4	17.6	72.5
Ours	**42.5**	**13.0**	**77.8**	**35.0**	**11.5**	**56.3**	**52.7**	**16.8**	**73.5**

Table 2: Results of CHAIR hallucination evaluation for the open-ended caption generation task. Lower $CHAIR_s$, $CHAIR_i$ scores and higher Recall scores indicate better performance.

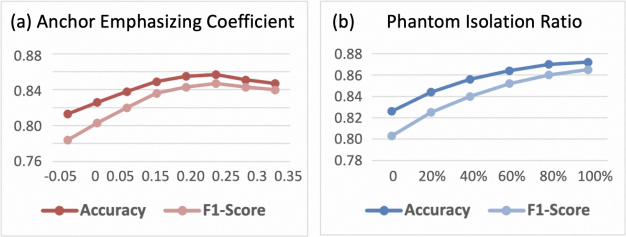

Figure 7: Ablation results of (a) anchor emphasizing coefficient α and (b) phantom isolation ratio.

2023), Qwen2.5-VL-7B (Wang et al. 2024), MiniGPT-4-7B (Chen et al. 2023a), and mPLUG-Owl2 (Ye et al. 2024).

5.3 Implementation Details

All experiments are conducted on NVIDIA A6000 GPUs with 48GB VRAM. We apply TAF as a plug-in module during inference without modifying the model's pre-trained parameters. The anchor emphasizing coefficient α is set to 0.2 unless otherwise specified. Our TAF remains stable across $\lambda_{\mathcal{A}}$, $\lambda_{\mathcal{P}}$, and other hyperparameters, with evidence in the Appendix. All evaluations are conducted using five random seeds, and results are reported as averages across runs.

5.4 Comparison with Existing Methods

POPE Evaluation. We compared the performance of various existing methods and our TAF under different settings, namely *Random*, *Popular*, and *Adversarial*, on the offline POPE benchmark (Li et al. 2023b; Chen et al. 2024d). The results, as shown in Table 1, demonstrate that our TAF consistently outperforms all baseline methods across all settings. These results underscore the effectiveness of our TAF in handling diverse conditions.

CHIAR Evaluation. We evaluated the effectiveness of our model on open-ended caption generation using the CHAIR benchmark (Rohrbach et al. 2018). Specifically, we employed the input prompt "Please describe this image in detail." The evaluation metrics included instance-level $CHAIR_i$, sentence-level $CHAIR_s$, and Recall, with experimental results presented in Table 2. It is evident that our TAF achieves competitive performance across all three evaluation metrics. These results demonstrate the robustness of our TAF in generating accurate image captions across different evaluation levels.

5.5 Ablation Study

We conducted ablation studies on the POPE benchmark using LLaVA-1.5-7B as a representative model.

Anchor Emphasizing Coefficient. We evaluated the effect of varying the Anchor Emphasizing Coefficient α. This ratio controls the extent to which anchor tokens are amplified within the attention mechanism during the filtering process. As shown in Figure 7(a), increasing the Anchor Emphasizing Coefficient enhances hallucination suppression, peaking at $\alpha = 0.2$. Further increases yield diminishing returns.

Phantom Isolation Ratio. The Phantom Isolation Ratio controls the suppression strength applied to the T2V attention logits of phantom tokens in the visual-active layers. When the ratio is 0, no isolation occurs, while 100% corresponds to complete isolation, where phantom tokens have no influence on the visual modality in relevant layers. As shown in Figure 7(b), higher isolation ratios greatly improve hallucination mitigation, with the best performance achieved at a ratio of 100%, corresponding to complete isolation.

5.6 Effectiveness of LFM Components

Figure 6 illustrates how LFM mitigates hallucinations through its two key components. In (a), raising α corrects the table's shape from *"rectangle"* to *"oval,"* reinforcing visual grounding. In (b), increasing isolation ratio, revising the chair count from *"two"* to *"three."* Anchor emphasizing and phantom isolation thus jointly enhance robustness.

6 Conclusion

In this work, we conduct a token-level analysis to investigate the potential causes of hallucinations in LVLMs, identifying disproportionate influence from certain textual tokens and insufficient utilization of critical visual cues. To address these issues, we propose a training-free, plug-and-play hallucination mitigation strategy, Token-Asymmetric Filtering (TAF). TAF focuses on the visual-active layers, selectively filtering unreliable text-to-visual signals and enhancing critical visual cues. Experimental results demonstrate that TAF effectively mitigates hallucinations, significantly improving the alignment between textual and visual modalities without the need for additional training.

Acknowledgments

This project is supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (Award Number: MOE-T2EP20122-0006).

References

Agarwal, V.; Shetty, R.; and Fritz, M. 2020. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9690–9698.

Agrawal, A.; Batra, D.; and Parikh, D. 2016. Analyzing the behavior of visual question answering models. *arXiv preprint arXiv:1606.07356*.

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.

An, W.; Tian, F.; Leng, S.; Nie, J.; Lin, H.; Wang, Q.; Dai, G.; Chen, P.; and Lu, S. 2024. Agla: Mitigating object hallucinations in large vision-language models with assembly of global and local attention. *arXiv preprint arXiv:2406.12718*.

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.

Biten, A. F.; Gómez, L.; and Karatzas, D. 2022. Let there be a clock on the beach: Reducing object hallucination in image captioning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1381–1390.

Chen, J.; Zhang, T.; Huang, S.; Niu, Y.; Zhang, L.; Wen, L.; and Hu, X. 2024a. ICT: Image-Object Cross-Level Trusted Intervention for Mitigating Object Hallucination in Large Vision-Language Models. *arXiv preprint arXiv:2411.15268*.

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 preprint arXiv:2306.15195*.

Chen, L.; Sinavski, O.; Hünermann, J.; Karnsund, A.; Willmott, A. J.; Birch, D.; Maund, D.; and Shotton, J. 2024b. Driving with llms: Fusing object-level vector modality for explainable autonomous driving. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 14093–14100. IEEE.

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

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

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*.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. arXiv:2305.06500.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

Fu, C.; Chen, P.; Shen, Y.; Qin, Y.; Zhang, M.; Lin, X.; Yang, J.; Zheng, X.; Li, K.; Sun, X.; Wu, Y.; and Ji, R. 2024. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. arXiv:2306.13394.

Gong, X.; Ming, T.; Wang, X.; and Wei, Z. 2024. Damro: Dive into the attention mechanism of lvlm to reduce object hallucination. *arXiv preprint arXiv:2410.04514*.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hu, M.; Pan, S.; Li, Y.; and Yang, X. 2023. Advancing medical imaging with language models: A journey from n-grams to chatgpt. *arXiv preprint arXiv:2304.04920*.

Huang, Q.; Dong, X.; Zhang, P.; Wang, B.; He, C.; Wang, J.; Lin, D.; Zhang, W.; and Yu, N. 2024a. 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.

Huang, W.; Liu, H.; Guo, M.; and Gong, N. Z. 2024b. Visual hallucinations of multi-modal large language models. *arXiv preprint arXiv:2402.14683*.

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*.

Jain, J.; Yang, J.; and Shi, H. 2024. Vcoder: Versatile vision encoders for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27992–28002.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12): 1–38.

Jiang, Z.; Chen, J.; Zhu, B.; Luo, T.; Shen, Y.; and Yang, X. 2024. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. *arXiv preprint arXiv:2411.16724*.

Lee, J.; Wang, Y.; Li, J.; and Zhang, M. 2024. Multimodal reasoning with multimodal knowledge graph. *arXiv preprint arXiv:2406.02030*.

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, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Li, L.; Xie, Z.; Li, M.; Chen, S.; Wang, P.; Chen, L.; Yang, Y.; Wang, B.; Kong, L.; and Liu, Q. 2024. VLFeedback: A Large-Scale

AI Feedback Dataset for Large Vision-Language Models Alignment. *arXiv preprint arXiv:2410.09421*.

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*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, 740–755. Springer.

Liu, F.; Lin, K.; Li, L.; Wang, J.; Yacoob, Y.; and Wang, L. 2023a. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024. 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. 2023b. Visual Instruction Tuning.

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

Liu, H.; Zhu, Y.; Kato, K.; Kondo, I.; Aoyama, T.; and Hasegawa, Y. 2023d. Llm-based human-robot collaboration framework for manipulation tasks. *arXiv preprint arXiv:2308.14972*.

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.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lyu, X.; Chen, B.; Gao, L.; Song, J.; and Shen, H. T. 2024. Alleviating hallucinations in large vision-language models through hallucination-induced optimization. *arXiv preprint arXiv:2405.15356*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Pi, R.; Han, T.; Xiong, W.; Zhang, J.; Liu, R.; Pan, R.; and Zhang, T. 2024. Strengthening multimodal large language model with bootstrapped preference optimization. In *European Conference on Computer Vision*, 382–398. Springer.

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

Sun, Z.; Shen, S.; Cao, S.; Liu, H.; Li, C.; Shen, Y.; Gan, C.; Gui, L.-Y.; Wang, Y.-X.; Yang, Y.; et al. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

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

Wang, S.; Zhao, Z.; Ouyang, X.; Wang, Q.; and Shen, D. 2023. Chatcad: Interactive computer-aided diagnosis on medical image using large language models. *arXiv preprint arXiv:2302.07257*.

Weijia, S.; Sewon, M.; Michihiro, Y.; Minjoon, S.; Rich, J.; Mike, L.; and Wen-tau, Y. 2023. REPLUG: Retrieval-augmented black-box language models. *arXiv preprint ArXiv:2301.12652*.

Woo, S.; Kim, D.; Jang, J.; Choi, Y.; and Kim, C. 2024. Don't Miss the Forest for the Trees: Attentional Vision Calibration for Large Vision Language Models. *arXiv preprint arXiv:2405.17820*.

Wu, J.; Liu, Q.; Wang, D.; Zhang, J.; Wu, S.; Wang, L.; and Tan, T. 2024. Logical closed loop: Uncovering object hallucinations in large vision-language models. *arXiv preprint arXiv:2402.11622*.

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.

Xue, L.; Shu, M.; Awadalla, A.; Wang, J.; Yan, A.; Purushwalkam, S.; Zhou, H.; Prabhu, V.; Dai, Y.; Ryoo, M. S.; et al. 2024. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*.

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.

Yu, T.; Yao, Y.; Zhang, H.; He, T.; Han, Y.; Cui, G.; Hu, J.; Liu, Z.; Zheng, H.-T.; Sun, M.; et al. 2024. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13807–13816.

Zhao, Z.; Wang, B.; Ouyang, L.; Dong, X.; Wang, J.; and He, C. 2023. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*.

Zhong, W.; Feng, X.; Zhao, L.; Li, Q.; Huang, L.; Gu, Y.; Ma, W.; Xu, Y.; and Qin, B. 2024. Investigating and mitigating the multimodal hallucination snowballing in large vision-language models. *arXiv preprint arXiv:2407.00569*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.

Zhou, Y.; Cui, C.; Rafailov, R.; Finn, C.; and Yao, H. 2024. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*.

Zhu, D.; Chen, J.; Shen, X.; xiang Li; and Elhoseiny, M. 2023. MiniGPT-4: Enhancing Vision-language Understanding with Advanced Large Language Models.