This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Same Attention, Different Truths: Put Logit-Lens over Visual Attention to Detect and Mitigate LVLM Object Hallucination

Zichuan Wang[1,2] Songlin Yang[3] Bo Peng[2*] Zhenchen Tang[1,2] Yang Li[2]
Beibei Dong[2] Jing Dong[2]
[1]School of Artificial Intelligence, University of Chinese Academy of Sciences
[2]New Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Science
[3]Hong Kong University of Science and Technology
{wangzichuan2024,tangzhenchen2024,liyang2022,dongbeibei2022}@ia.ac.cn
syangds@connect.ust.hk,{bo.peng,jdong}@nlpr.ia.ac.cn

Abstract

Large Vision-Language Models (LVLMs) often suffer from object hallucination, generating objects that are absent from the image. Prior work largely attributes this to insufficient visual attention. However, we find that both real and hallucinated objects receive equally strong visual attention in model's mid-to-late layers, suggesting that the key issue may not be how much the model attends, but ***what it attends to and why****. To this end, we decode the visual features of high-attention regions using Logit Lens, and observe that regions corresponding to real objects can be correctly decoded to the target object tokens, whereas those for hallucinated objects cannot. Building on this, we identify two hallucination mechanisms:* ***(i) visual uncertainty****, triggered by semantically similar or confusable regions, masking these regions eliminates the hallucination.* ***(ii) contextual prior****, triggered by strong co-occurrence priors, even when the initially attended region is masked, the hallucination persists and attention drifts to other regions. Based on these findings, we propose a simple yet effective training-free* ***Detect–Mitigate framework*** *comprising a Logit-Lens Consistency Check to detect hallucination and targeted remedies: High-Attention Regions Masking (HARM) for visual uncertainty hallucination, and Visual Evidence Enhanced Decoding (VEED) for contextual prior hallucination. Our approach achieves state-of-the-art results on multiple hallucination benchmarks. Code will be available.*[1]

1. Introduction

Large Vision-Language Models (LVLMs) have made significant strides in recent years, demonstrating impressive capabilities on cross-modal tasks [3, 8, 10, 19, 20, 25]. However, they still suffer from object hallucination [4, 22, 27], i.e., generating objects that are not present in the image, undermining their reliability and practical use.

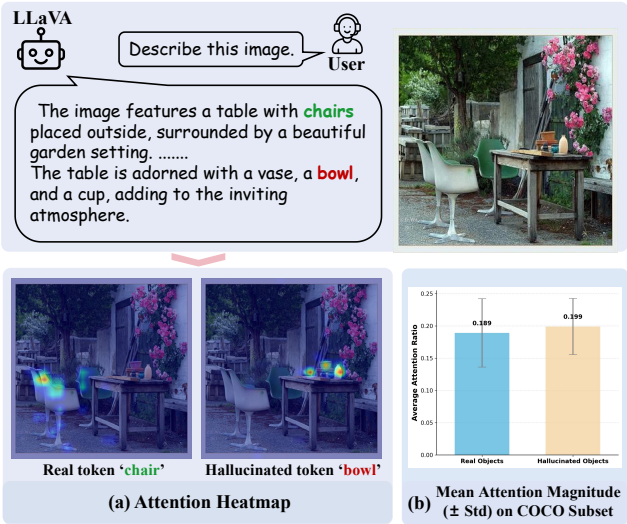

Figure 1. Different from prior work that attributes object hallucination to insufficient visual attention, we find that real and hallucinated objects often have the same attention magnitudes, as shown qualitatively in **(a)** and quantitatively in **(b)**, reflecting "Same Attention, Different Truths".

A prevailing view attributes hallucination to insufficient visual attention, either because textual priors dominate the multimodal interaction and suppress visual focus [5, 16, 22, 36, 40], or because improper processing of visual information leads to inadequate attention to salient regions [2, 15, 17]. Accordingly, many methods attempt to amplify or redistribute attention [11, 15, 17, 28, 43], or inject extra visual guidance during training/inference [7, 18,

*Corresponding author.
[1]https://github.com/wzczc/SADT

34, 45]. However, our experiments reveal that this explanation is incomplete: *the key issue may not be how much the model attends to the image, but what it attends to and why*. To support this claim, we first examine **how much** the model attends. By quantifying the attention strength associated with object tokens, we find that visual attention consistently peaks in the mid-to-late layers, which we define as the "image attention stage". Both real and hallucinated object tokens exhibit similar attention magnitudes and focus on specific regions in this stage, as shown in Fig. 1. Thus, the occurrence of hallucination cannot be explained simply by insufficient visual attention.

Next, to understand **what** the model attends to, we analyze the semantics behind these high-attention regions using the Logit Lens [31], which uses the model's textual decoding head to translate visual representations into text tokens. We observe that regions corresponding to real object tokens can be accurately decoded into their target tokens, whereas those corresponding to hallucinated object tokens fail to produce semantically consistent outputs. This reveals a semantic misalignment between visual semantics and output tokens within high-attention regions, showing that hallucination is typically accompanied by inconsistent semantic decoding, rather than by the lack of visual attention.

To further probe **why** the model attends to these regions when hallucination occurs, we mask these regions and examine how the model's output changes. This intervention reveals two distinct hallucination mechanisms. **(1) Visual Uncertainty Hallucination**: triggered by confusing or semantically related regions (e.g., a fuzzy ceramic-like area mistaken for a "bowl"). Masking these high-attention regions directly eliminates the hallucination. This shows that the model cannot extract precise semantics from highly uncertain visual evidence, leading to decoding inconsistency. **(2) Contextual Prior Hallucination**: dominated by co-occurrence priors (e.g., generating "microwave" in a description of a kitchen even though it is absent). In this case, masking the high-attention regions does not remove the hallucination, and attention shifts to other regions. Although these regions can be correctly decoded into their target tokens, they are not selected in the final output due to the suppression by strong priors. Here, the attention mechanism behaves more like a "routine": the model has learned that it must attend to some region before generating an object and thus anchors attention merely to satisfy this requirement.

Based on these findings, we propose a simple yet effective training-free **Detect–Mitigate framework** to precisely locate and reduce hallucinations. First, we propose the Logit-Lens Consistency Check to detect hallucinated tokens by comparing whether the semantic decoding of high-attention regions in the mid-to-late layers is consistent with the token being generated. Then, we propose corresponding mitigation methods for each category. For the visual uncertainty cases, we perform simple High-Attention Regions Masking (HARM) to remove unreliable visual evidence. For the contextual prior cases, we propose Visual Evidence Enhanced Decoding (VEED). Here, the high-attention regions can still be decoded to the correct tokens via Logit Lens, but are suppressed during generation due to the strong priors. We therefore inject the true visual evidence from these regions into the generation process, enhancing the contribution of image-grounded semantics and suppressing contextual priors. Finally, we apply our strategy to multiple LVLMs on several hallucination benchmarks, achieving SOTA results, demonstrating its effectiveness and generality. In summary, our contributions are as follows:

- We reveal the "same attention, different truth" phenomenon, showing the key of object hallucination is the semantic inconsistency between high-attention visual features and the text output, not the attention strength. Based on this, we categorize object hallucination into two core mechanisms: visual uncertainty and contextual prior.
- We propose a simple yet effective training-free "Detect-Mitigate" framework, which includes: (1) the Logit-Lens Consistency Check for detection and (2) two mitigation strategies targeting each mechanism: High-Attention Regions Masking and Visual Evidence Enhanced Decoding.
- Our method achieves SOTA performance on multiple object hallucination benchmarks across several LVLMs, proving its effectiveness and generalizability.

2. Related Work

2.1. Large Vision-Language Models

Recent years have witnessed remarkable progress in Large Vision-Language Models (LVLMs), which integrate language models with visual encoders to perform unified multimodal capacity. Early work such as CLIP [32] established large-scale vision–language pretraining via contrastive image–text alignment. Subsequent models, including BLIP-2 [19] and Flamingo [1], further bridged modality gaps via learnable adapters or cross-attention mechanisms. More recently, instruction-tuned LVLMs such as LLaVA [25, 26], Shikra [6], mPLUG-Owl2 [42], Qwen-VL [3] and InternVL [8] have demonstrated impressive open-ended generation and reasoning capabilities. Despite these advances, LVLMs still frequently hallucinate objects, i.e., describe objects that are absent from the image, which undermines their reliability in practical applications.

2.2. Object Hallucination in LVLMs

Object hallucination is a critical challenge for LVLMs. Existing mitigation methods fall into two categories. The first is training-stage methods, such as introducing additional datasets [7, 24], incorporating new training objectives [13, 29], and reinforcement learning methods [41, 44, 46]. How-

ever, they are costly in terms of time and resources. The second is inference-stage methods, including contrastive decoding [18, 21], guided decoding [9, 45], and visual enhancement [2, 43]. Recently, a growing body of work has begun to investigate the internal mechanisms of hallucination at inference time, with many studies attributing it to insufficient visual attention. [18, 28, 40, 43] find that textual priors often dominate multimodal interaction. In parallel, [2, 15, 17, 36] show that improper processing of visual information also results in weak visual attention. Building on these, [15, 17, 43] propose reallocating or amplifying attention to strengthen the visual pathway, while [18, 28, 37] leverage visually guided to counteract language priors. Our work follows the attention-based analysis but takes a step further. We argue that the "insufficient visual attention" explanation is incomplete, as our findings show real and hallucinated objects can receive comparable attention magnitudes, while the semantic consistency between the attended regions and the textual output may be the key.

3. Object Hallucination Analysis

In this section, we analyze the mechanisms underlying object hallucination in LVLMs. We begin with the necessary preliminaries, then compare the attention patterns of hallucinated and non-hallucinated objects and use the Logit Lens to identify their semantic differences. Finally, we conduct intervention experiments that categorize object hallucination into two distinct types.

3.1. Preliminary

LVLM Generation Process. LVLMs typically consists of a vision encoder, a cross-modal projector and a LLM. Given an image I and a textual prefix $Q_{<t} = \{q_1, \ldots, q_{t-1}\}$, the vision encoder maps I to image tokens $V = \{v_1, \ldots, v_N\}$, which the projector M transforms into $\tilde{v}_i = M(v_i)$. These tokens are then concatenated with system tokens $S = \{s_1, \ldots, s_M\}$ and the text prefix to form the input sequence $X = [s_1, \ldots, s_M, \tilde{v}_1, \ldots, \tilde{v}_N, q_1, \ldots, q_{t-1}]$, which is fed into a multi-layer transformer decoder with causal masking. At generation step t, the output head W_U maps the hidden state $h_t^{(L)}$ of the last layer L to the next-token distribution:

$$p(x_t \mid I, X_{<t}) = \text{softmax}(W_U h_t^{(L)}). \tag{1}$$

Attention Mechanism. In LVLMs, attention transfers information between tokens. The autoregressive transformer uses causal attention: at step t, the query can only attend to tokens at positions $j \leq t$, while positions $j > t$ are masked. Let q_t be the query, and k_j be the keys. The mask m_{tj} is defined as $m_{tj} = 0$ when $j \leq t$ and $-\infty$ otherwise. The attention weight between tokens are then computed as:

$$\alpha_{tj} = \text{softmax}\left(\frac{q_t^\top k_j}{\sqrt{d_k}} + m_{tj}\right). \tag{2}$$

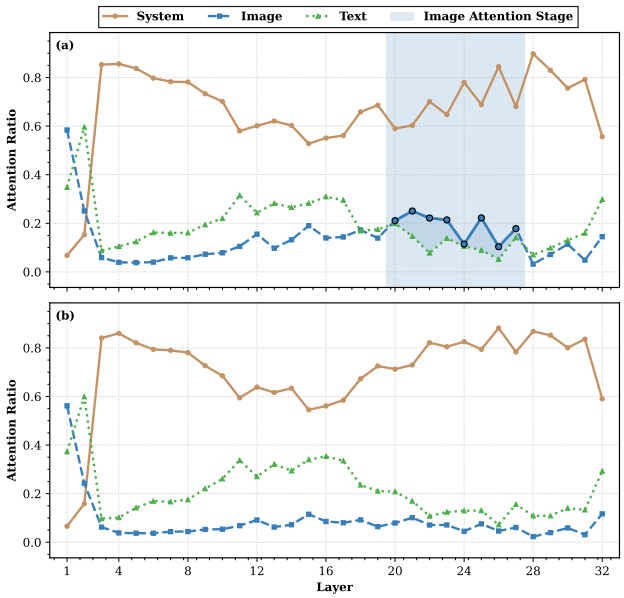

Figure 2. Attention Ratio across layers. **(a)** Object attention; **(b)** Non-object attention. Non-object tokens exhibit minimal attention to the image, while for object tokens we partition the attention into four stages according to their variations through layers, with the mid-to-late layers (20-27) being the Image-Attention Stage.

Logit Lens. Logit Lens is an interpretability method that decodes the hidden states of intermediate layers in LLM into vocabulary space, revealing how information evolves across layers. Given a hidden state $h_t^{(l)}$, the output head W_U produces a probability distribution over the vocabulary:

$$p_{\text{lens}}^{(l)}(\cdot \mid h_t^{(l)}) = \text{softmax}\left(W_U\, h_t^{(l)}\right). \tag{3}$$

In this work, we apply it to the hidden states of image tokens to interpret how the model processes visual information at different layers.

3.2. Same Attention: Attention Magnitude Is Not to Blame

To analyze whether attention magnitude play a role in generating hallucinated object tokens, we adopt LLaVA-1.5-7B [26] as an example, randomly sampling 500 images from the COCO 2014 validation set [23] and use the prompt "Describe this image." for generation. For each generated token, we collect its attention distribution data across all the layers. We first compare the attention distributions of generated object tokens versus non-object tokens, as shown in Fig. 2. It is observable that, apart from the initial 1-2 layers, non-object tokens exhibit minimal attention to the image. In contrast, object tokens exhibit a four-stage pattern:

- Initialization Stage: In the first 1–2 layers, attention is high on both image and text, as the model rapidly aggregates multimodal context.

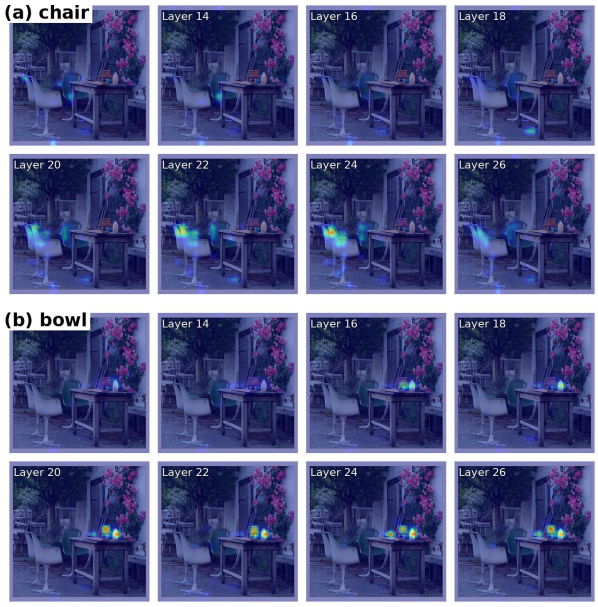

Figure 3. Both the real object token **(a)** 'chair' and the hallucinated object token **(b)** 'bowl' exhibit attention that converges on specific regions during the image-attention stage (layers 20–27).

- Text-Attention Stage: In the early-to-mid layers (3-19), text attention dominates while image attention remains relatively low.
- **Image-Attention Stage**: In the mid-to-late layers (20-27), attention to the image increases and becomes dominant, while text attention drops. The model focuses on extracting evidence related to the candidate objects.
- Language-Organization Stage: In the final layers (28-32), image attention decreases again while text attention rises, as the model is shifting to organize and generate a grammatically and logically coherent output.

We then ask whether real and hallucinated object tokens differ in how much they attend to the image. We employ CHAIR [33] to identify real and hallucinated object tokens and quantify their average image attention during the Image-Attention Stage, as shown in Fig. 1 (b). The results show no systematic difference between the two categories. To provide further qualitative evidence, we visualize the evolution of image attention across layers when generating object tokens. As shown in Fig. 3, both real and hallucinated object tokens focus their attention on specific regions during the Image-Attention Stage, with comparable magnitudes. These observations indicate that **object hallucination is not caused by insufficient visual attention.**

3.3. Different Truths: Semantic Inconsistency Revealed by Logit-Lens

The previous section showed that hallucination is not related to **how much** the model attends to the image. In the following, we show its relation to **what** the model attends to. Specifically, we use Logit-Lens [31] to "read out" the semantics of the attended visual regions. During the Image-Attention Stage, we project the hidden states of the top-k attended image tokens into the vocabulary space, thereby decoding what the model semantically "recognizes". The results are shown in Fig. 4. We find:

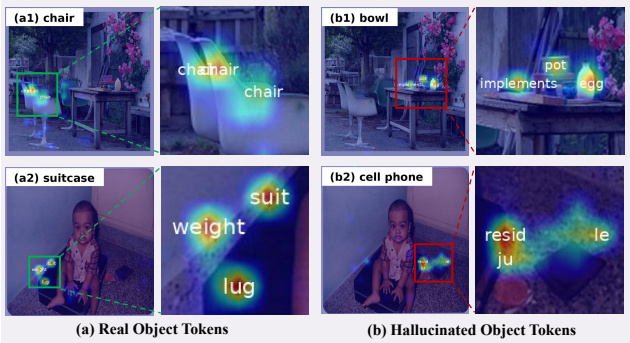

Figure 4. Via the Logit-Lens, semantics decoded from the high-attention regions in the Image-Attention Stage align with the **(a)** real object tokens, but diverge from **(b)** the hallucinated object tokens.

- For **Real Objects (Semantic Consistency)**: When the model generates a real object token (e.g., 'suitcase'), the highly-attended visual regions (i.e., the actual suitcase regions) can also be decoded to consistent tokens like 'suit' or 'lug' by Logit-Lens. This indicates that the real token attention is faithfully grounded in the visual evidence.
- For **Hallucinated Objects (Semantic Inconsistency)**: When the model generates a hallucinated object (e.g., 'cell phone'), a clear semantic inconsistency emerges. The highly-attended visual regions (e.g., a 'blurry patch on the ground') do not decode to the hallucinated target token by Logit-Lens. This reveals a critical disconnect: the semantics "read" from the attended region are inconsistent with the final output.

This finding uncovers a key phenomenon of object hallucination: **hallucinated object tokens exhibit a semantic inconsistency, where the visual evidence extracted by the attention mechanism does not support the generated token.** This insight directly motivates our subsequent detection method, and empirical evidence for this phenomenon can be seen in the detection experiments in Sec. 5.1.

3.4. Two Types: Visual Uncertainty and Contextual Prior

The previous section showed that hallucinations can appear around the regions with strong attention during the Image-Attention Stage. In this section, we further investigate **why** the model attends to these regions by devising a High-Attention Regions Masking experiment.

Let the attention weight at layer l on the j-th image token during the t-th generation step (targeting object token o_t) be $\alpha_{tj}^{(l)}$. Let $\mathcal{S}_{\text{IA}}$ be the set of layers in the Image-Attention Stage. We first obtain the average attention that o_t assigns to each image token j in $\mathcal{S}_{\text{IA}}$:

$$\bar{\alpha}_{tj} = \frac{1}{|\mathcal{S}_{\text{IA}}|} \sum_{l \in \mathcal{S}_{\text{IA}}} \alpha_{tj}^{(l)}. \tag{4}$$

We then select the top-k high-attention image tokens $\Omega_t = \text{TopK}(\bar{\alpha}_t, k)$, mask corresponding patches in the image and regenerate the response to observe the model's behavior. As illustrated in Fig. 5, this experiment reveals two types of object hallucinations with distinct underlying causes:

- **Visual Uncertainty Hallucination**: In these cases, the hallucinated token is directly related with high-attention regions. When we mask these regions, the hallucinated token disappears, indicating that the hallucination is visually rooted. Empirically, such regions are often ambiguous, blurry, or uncertain areas (e.g., a round, ceramic-like area leading to the generation of "bowl"; or a dark, blurry rectangular region resulting in the output "cell phone"). These observations suggest that the model may attempt to forcibly extract certain semantics from highly uncertain visual evidence, which is insufficient to support the generated object, leading to semantic inconsistency. Once this uncertain evidence source is removed, the model loses its visual anchor for generating hallucination.
- **Contextual Prior Hallucination**: In contrast, we also observe hallucination where masking the high-attention regions does not remove the hallucinated token. Instead, hallucination persists and attention shifts to other regions. Logit Lens analysis further shows that these attended regions can be decoded into the correct objects actually present in the image, yet the model still outputs the hallucinated token. This pattern is consistent with the generation being dominated by strong contextual priors [28] (e.g., outputting "microwave" in a kitchen description even when no microwave is present). In such cases, the attention mechanism is more like a procedural "routine": its internal rules require it to "attend somewhere" before generating an object. So it arbitrarily "grounds" its attention on some regions to satisfy this procedural requirement, while the actual generation decision is mainly controlled by the contextual prior.

We further perform a statistical analysis on the hallucination instances over 500 images. For LLaVA-1.5-7b, the ratio of "Visual Uncertainty" to "Contextual Prior" hallucination is approximately 2:1. The findings in this section expose two different types of object hallucination, demonstrating that a single, one-size-fits-all mitigation strategy is likely insufficient, thereby laying a solid foundation for our subsequent design of a targeted mitigating method.

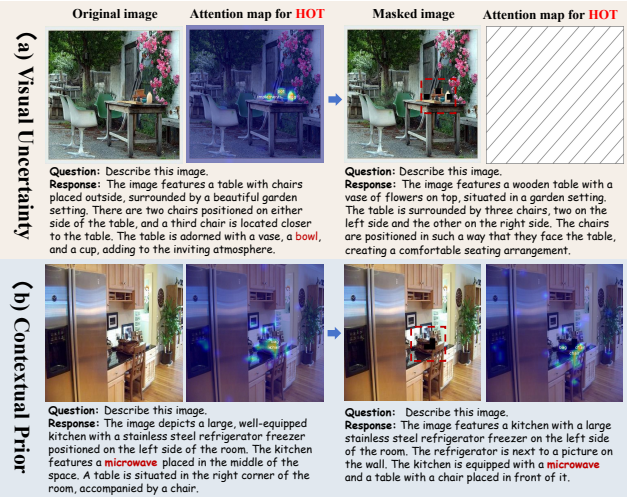

Figure 5. Two types of hallucination. **(a)** Disappears after masking high-attention regions, driven by visual uncertainty. **(b)** Persists after masking with attention shifting to other regions, driven by contextual prior. HOT here means Hallucinated Object Token.

4. Detect-Mitigate Framework For Object Hallucination

Based on the analysis in Sec. 3, we propose a simple yet effective Detect–Mitigate framework, as shown in Fig. 6. We first introduce a Logit-Lens Consistency Check (LLCC) to detect potential hallucinated tokens in real time (motivated by Sec. 3.3). We then determine which type (from Sec. 3.4) each detected hallucination belongs to and apply a targeted mitigation strategy. Details are as follows.

4.1. Hallucination Detection via Logit-Lens Consistency Check

Our detection module aims to identify object hallucination in real-time within the generated sequence. This process is based on the finding from Sec. 3.3: real objects are semantically consistent with their visual evidence, whereas hallucinated objects are not. The detection process is divided into three stages: (1) Object Token Identification, (2) Semantic Consistency Check and (3) Hallucination Classification.

4.1.1. Object Token Identification

To precisely locate object tokens while avoiding unnecessary computation on non-object tokens, we design an identification step. First, inspired by the observation from Sec. 3.2 that object tokens exhibit significantly higher image attention than non-object tokens, we calculate the average image attention $\mathcal{A}_{img}(o_t)$ for each generated token o_t within the Image-Attention Stage $\mathcal{S}_{\text{IA}}$:

$$\mathcal{A}_{img}(o_t) = \frac{1}{|\mathcal{S}_{\text{IA}}|} \sum_{l \in \mathcal{S}_{\text{IA}}} \left(\sum_{j=1}^{N_{img}} \alpha_{tj}^{(l)} \right), \tag{5}$$

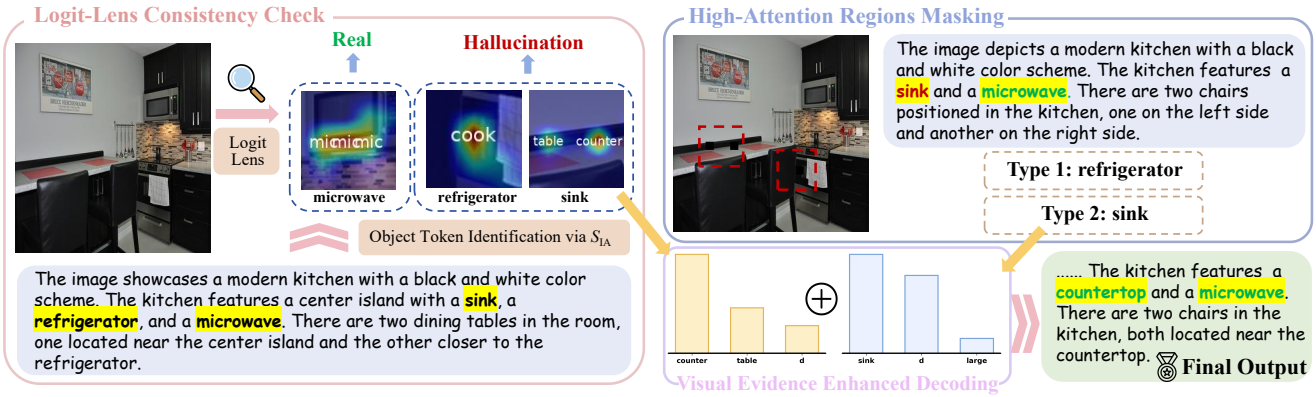

Figure 6. The overall structure of our Detect-Mitigate framework. It consists of three components. First, **Logit-Lens Consistency Check (LLCC)** detects hallucinated object tokens in the response. Then, **High-Attention Regions Masking (HARM)** categorizes them into two different types of hallucination while generating an intermediate output. If type 2 hallucination is detected, the output is further processed by **Visual Evidence Enhanced Decoding (VEED)** to produce the final result.

where $\alpha_{tj}^{(l)}$ is the attention weight of o_t on the j-th image token at layer l, and N_{img} is the total number of image tokens. We treat o_t with $\mathcal{A}_{\text{img}}(o_t) > \tau_{\text{attn}}$ as object tokens, where τ_{attn} is a threshold hyper-parameter.

4.1.2. Semantic Consistency Check

For each object token, we first collect its attention to all image tokens within $\mathcal{S}_{\text{IA}}$ and select the Top-k attended tokens:

$$\Omega_t = \text{TopK}\left(\alpha_{tj}^{(l)}, k\right), l \in \mathcal{S}_{\text{IA}}. \tag{6}$$

We apply Logit Lens to each hidden state $h(\omega_t^i), \omega_t^i \in \Omega_t$ and take the highest-probability decoded token:

$$v_t^i = \arg\max_{v \in \mathcal{V}} \left[\text{softmax}\left(W_U h(\omega_t^i)\right)\right]_v, i = 1, \ldots, k. \tag{7}$$

where $\mathcal{V}$ denotes the vocabulary. Then we compare whether o_t is semantically consistent with each token v_t^i. Since v_t^i might be a synonym of o_t (e.g., o_t=car, v_t^i=vehicle), we adopt a standard similar to [33, 38], using WordNet [30] or another semantic similarity function $\text{Sim}(\cdot, \cdot)$:

$$\text{Label}(o_t) = \begin{cases} 0 \text{ (Real)}, & \text{if } \text{Sim}(o_t, v_t^i) > \tau_{\text{sim}}, \\ 1 \text{ (Hallucination)}, & \text{otherwise.} \end{cases} \tag{8}$$

where τ_{sim} is the semantic similarity threshold.

4.1.3. Hallucination Classification

When a token o_t is detected as "hallucination", we further classify its type. Following the same procedure as Sec. 3.4, we mask the high-attention regions within $\mathcal{S}_{\text{IA}}$ and observe whether the hallucination "disappears". Let the newly generated output sequence be O_{new}, then:

$$\text{Type}(o_t) = \begin{cases} 1 \text{ (Visual Uncertainty)} & \text{if } o_t \notin O_{\text{new}}, \\ 2 \text{ (Contextual Prior)} & \text{if } o_t \in O_{\text{new}}. \end{cases} \tag{9}$$

4.2. Targeted Mitigation Methods for Hallucination

A single strategy cannot effectively handle hallucinations arising from different mechanisms. We therefore propose two simple yet effective methods tailored to each class.

4.2.1. Visual Uncertainty: High-Attention Region Masking

As analyzed in Sec. 3.4, this type of hallucination occurs when the model anchors its attention on a high-uncertainty visual region, and the generation heavily depends on this erroneous evidence. Therefore, we employ a simple masking strategy, identical to that in Sec. 3.4.

Let all the Top-k high-attention image tokens of each hallucinated object token be Ω and the original image be I. Its binary mask M satisfies:

$$\mathbf{M} = \begin{cases} 1, & \text{if pixel/patch } p \in \Omega \\ 0, & \text{otherwise} \end{cases}. \tag{10}$$

Let μ be a simple replacement value, e.g., the mean image color or a zero vector (black). The masked image I^{mask} is:

$$I^{\text{mask}} = (1 - \mathbf{M}) \odot I + \mathbf{M} \odot \mu. \tag{11}$$

We then regenerate with the same prompt using I^{mask}. By removing the high-uncertainty evidence, this strategy effectively reduces hallucination of this type.

4.2.2. Contextual Prior: Visual Evidence Enhanced Decoding

The second type is insensitive to masking. Its root cause is that strong contextual priors dominate the output, while the real visual evidence is ignored. To address this, we propose Visual Evidence Enhanced Decoding (VEED). During the decoding stage of the language model, we directly inject visual semantics from high-attention regions to increase their

weight in the final logits, thereby exploiting correct visual evidence and suppressing the influence of the priors.

Let $h(\omega_t^{\max})$ denote the hidden state of the most attended visual region at step t. We extract its visual logits using Logit Lens (computed in Sec. 4.1): $z_t^{\text{vis}} = W_U h(\omega_t^{max})$. Let $z_t^{\text{mask}} = \text{logits}_\theta(y_t \mid y_{<t}, I^{\text{mask}})$ be the model's decoding logits at step t when the high-attention region is masked. We fuse them to form the enhanced distribution for type 2:

$$\begin{aligned} p_t &= \text{softmax}\left((1-\alpha)\cdot z_t^{\text{vis}} + \alpha \cdot z_t^{\text{mask}}\right) \qquad (12) \\ &= \text{softmax}\Big((1-\alpha)\cdot W_U h(\omega_t^{\max}) \\ &\quad + \alpha \cdot \text{logits}_\theta(y_t \mid y_{<t}, I^{\text{mask}})\Big) \end{aligned}$$

Here, α controls the strength of visual-evidence injection: a lower value gives higher weight to visual semantics from the high-attention region in the final decision.

5. Experiments

In this section, we empirically validate the effectiveness of our proposed Detect–Mitigate framework. We first assess the hallucination detection method in Sec. 4.1, and then the mitigation method in Sec. 4.2.

5.1. Hallucination Detection

5.1.1. Experimental Settings

Following [12, 15], we randomly sample 500 images from COCO2014 dataset, generate descriptions with LLaVA-1.5-7B, and use CHAIR [33] to label hallucinated and real objects. We compare our approach against three recent methods: Uncertainty Score [47], which identifies hallucinations via uncertainty estimates; InterConf [14], which aggregates decoded probabilities from all layers; and SVAR [15], which exploits attention weights from intermediate layers 5–18. We set τ_{attn} to 0.15 and k to 3. Following [38], we set τ_{sim} to 0.8 to match semantically similar words. Additional parameter experiments are in the appendix.

5.1.2. Results and Analysis

As shown in Tab. 1, LLCC achieves the best Precision, Recall, and F1. By explicitly localizing object-token attention and performing a semantic consistency check on the attended regions, it directly tests whether the generation is supported by visual evidence. In contrast, baselines rely on decoding probabilities or cross-layer aggregates, weakening their causal link to the image. Uncertainty Score measures uncertainty based on probability, which lacks a theoretical link between probability and uncertainty and is undermined by LLM's overconfidence [35]. InterConf merely assesses whether the object exists in the image, lacking explicit localization, so language priors or misleading visual cues can make it wrong, and its reliance on all layers further degrades accuracy. SVAR focuses on attention "quantity" (Summed Ratio) over "quality". Our analysis (Sec. 3.2 / Sec. 3.3) shows that attention aligns with image evidence primarily in the later Image-Attention Stage, while it aggregates signals from non-image-dominant layers, which introduces noise and dilutes the signal. Our findings (Sec. 3.4) also suggest quantity maybe not a very reliable metric. While for our method, it uniquely assesses the "reasonableness of the generation" by first locking onto the visual source for object tokens and then checking if that source provides consistent semantic support. Notably, it can also identify different causal types of hallucinations, which prior methods cannot.

Table 1. Hallucination Detection Results. Best results are in **bold**, and second-best are underlined.

Method	Precision	Recall	F1 Score
Uncertainty Score	0.5965	0.6415	0.6182
InterConf	0.6717	0.6907	0.6811
SVAR	0.6500	0.7222	0.6842
LLCC(Ours)	**0.7870**	**0.7955**	**0.7932**

5.2. Hallucination Mitigation

5.2.1. Experimental Settings

Benchmark and metrics. We evaluate on two standard benchmarks: **CHAIR** [33] and **AMBER** [38]. **CHAIR** measures whether objects mentioned in the generated text appear in the image and reports sentence-level and instance-level metrics (CHAIR_S and CHAIR_I, lower is better):

$$\text{CHAIR}_I = \frac{|\text{hallucinated objects}|}{|\text{all objects mentioned}|}, \text{CHAIR}_S = \frac{|\text{sentences with hallucinated object}|}{|\text{all sentences}|}$$

Following common settings [12, 15, 28], we randomly sample 500 images from the COCO2014 validation set for CHAIR evaluation. **AMBER** is a multi-dimensional benchmark comprising 1004 images with detailed object annotations. We report four metrics under its generative setting: CHAIR (rate of hallucinated objects), Cover (object coverage), Hal (fraction of responses containing any hallucination), and Cog (tendency to produce cognitively tempting target objects), where higher is better for Cover and lower is better for the others.

Baselines. We compare against three common decoding strategies and five representative hallucination mitigation methods. Greedy Search selects the most probable next token at each step. Beam Search maintains multiple hypotheses and outputs the highest-probability sequence. Nucleus Sampling samples from tokens whose cumulative probability exceeds a threshold p. VCD [18] contrasts outputs generated from the original and a perturbed image to suppress language priors. OPERA [12] introduces dynamic penalties on over-confident decoding steps. DeCo [37] selects an intermediate "anchor" layer to correct final-layer logits. Devils [15] exploits mid-layer visual signals by aligning multi-

Table 2. CHAIR hallucination evaluation results on multiple LVLMs. Best results are in **bold**, and second-best are underlined. DeCo and Devils do not report analyses on Qwen2-VL so their experimental settings are unspecified. We mark them as "–".

Method	LLaVA-1.5-7B		LLaVA-1.5-13B		Shikra-7B		Qwen2-VL-7B	
	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	$CHAIR_S$	$\downarrow CHAIR_I \downarrow$
Greedy	49.8	20.4	47.8	19.8	58.4	22.2	31.4	12.7
Beam	50.0	20.1	48.2	18.9	59.7	23.1	30.8	11.9
Nucleus	59.0	25.0	54.2	21.7	58.0	22.0	33.3	14.0
VCD	56.3	22.9	50.3	18.9	52.4	19.8	32.1	13.2
OPERA	42.9	18.7	42.1	16.4	38.1	16.7	28.0	10.3
DeCo	37.3	16.5	38.3	15.7	40.2	16.9	-	-
Devils	32.1	13.7	35.4	14.1	32.8	13.1	-	-
PAI	29.8	13.2	33.2	13.5	37.9	15.0	24.7	8.6
Ours	**26.8**	**10.0**	**31.3**	**12.4**	**31.4**	**12.7**	**24.0**	**8.3**

Table 3. AMBER hallucination evaluation results. Best results are in **bold**, and second-best are underlined.

Method	LLaVA-1.5-7B				LLaVA-1.5-13B				Shikra-7B			
	$CHAIR \downarrow$	$Cover \uparrow$	$Hal \downarrow$	$Cog \downarrow$	$CHAIR \downarrow$	$Cover \uparrow$	$Hal \downarrow$	$Cog \downarrow$	$CHAIR \downarrow$	$Cover \uparrow$	$Hal \downarrow$	$Cog \downarrow$
Greedy	6.9	51.0	32.0	3.3	6.8	52.0	31.7	3.5	10.6	52.0	47.0	5.2
Beam	7.8	49.7	35.5	4.1	8.7	48.7	38.8	3.8	10.0	51.3	44.1	4.8
Nucleus	7.9	51.1	35.5	4.3	8.4	51.5	38.1	3.9	11.1	52.2	49.4	5.7
VCD	6.4	**52.1**	33.2	2.9	7.6	**53.2**	35.3	3.4	10.3	**53.7**	45.2	5.0
OPERA	6.8	47.3	28.3	2.7	6.9	49.0	31.5	3.3	8.2	51.6	39.8	3.9
DeCo	6.6	46.4	27.6	2.8	6.2	48.5	27.7	2.9	7.5	48.2	37.9	4.0
Devils	3.5	50.2	19.9	1.3	5.6	51.1	30.6	3.4	6.3	51.4	32.1	2.8
PAI	4.2	50.7	18.4	1.6	4.9	51.9	26.2	3.0	5.9	50.9	32.7	3.4
Ours	**2.8**	51.2	**14.7**	**1.2**	**4.0**	52.0	**24.1**	**2.5**	**5.3**	52.3	**30.3**	**2.0**

head attentions. And PAI [28] amplifies the attention of image tokens and contrasts its output with no-image logits.

Implementation details. We implement our method with greedy decoding as a representative case and test on three VLMs: LLaVA-1.5 [26], Shikra [6] and Qwen2-VL [39]. For Beam search, we set $n_{beam} = 3$. For Nucleus sampling, we set top-$p = 0.9$ and temperature $\tau = 0.7$. Following the original papers, we pair VCD with nucleus sampling, apply OPERA and DeCo on top of beam search, and combine Devils and PAI with greedy decoding. All experiments are conducted on a single NVIDIA A100 40GB GPU.

5.2.2. Results and Analysis

As shown in Tab. 2 and Tab. 3, our method consistently reduces hallucination while maintaining (even slightly improving) object coverage. On CHAIR, we achieve the largest drop among the four models. On AMBER, we obtain the lowest hallucination rates across models while keeping strong coverage, which is slightly higher than the original 51.0 and second only to VCD's 52.1 on LLaVA-1.5-7B. Notably, most baselines decrease on Cover rate, indicating that they sacrifice valid content to mitigate hallucination, whereas our approach achieves the lowest hallucination rates without eroding the original expression. We attribute the advantage to our two-stage framework: (i) explicit object-token localization and cause-aware hallucination detection, and (ii) targeted mitigation: masking only a minimal set of image tokens and applying decoding enhancement only for the second hallucination type, thereby limiting impact on original content. Although absolute scores vary with model scale and architecture, the relative ranking remains stable, demonstrating robustness of our method. Additional ablations are provided in the appendix.

6. Conclusion

In this work, we revisit object hallucination in LVLMs via the lens of attention and Logit Lens analysis. By analyzing object token generation process, we reveal a "same attention, different truths" phenomenon: both real and hallucinated objects exhibit similar attention patterns in the image-attention stage, yet only the former are semantically decodable from the attended regions via Logits-Lens. Building on this, we propose a training-free Detect–Mitigate framework that first detects hallucinated object tokens via Logit-Lens Consistency Check and then applies cause-aware High Attention Regions Masking for visual uncertainty hallucination and Visual Evidence Enhanced Decoding for contextual prior hallucination. Extensive experiments demonstrate the effectiveness and generality of our method.

7. Acknowledgments

This work is supported by the National Key Research and Development Program of China under Grant No. 2025ZD0123504, the National Natural Science Foundation of China (NSFC) under Grant 62272460 and Jiangsu Provincial Science and Technology Major Project under Grant BG2024042.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 2

[2] Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Ping Chen, Xiaoqin Zhang, and Shijian Lu. 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*, pages 29915–29926, 2025. 1, 3

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 1, 2

[4] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024. 1

[5] Junzhe Chen, Tianshu Zhang, Shiyu Huang, Yuwei Niu, Linfeng Zhang, Lijie Wen, and Xuming Hu. Ict: Image-object cross-level trusted intervention for mitigating object hallucination in large vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4209–4221, 2025. 1

[6] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 2, 8

[7] Zhiyang Chen, Yousong Zhu, Yufei Zhan, Zhaowen Li, Chaoyang Zhao, Jinqiao Wang, and Ming Tang. Mitigating hallucination in visual language models with visual supervision. *arXiv preprint arXiv:2311.16479*, 2023. 1, 2

[8] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 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*, pages 24185–24198, 2024. 1, 2

[9] Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. Halc: Object hallucination reduction via adaptive focal-contrast decoding. *arXiv preprint arXiv:2403.00425*, 2024. 3

[10] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023. 1

[11] Xuan Gong, Tianshi Ming, Xinpeng Wang, and Zhihua Wei. Damro: Dive into the attention mechanism of lvlm to reduce object hallucination. *arXiv preprint arXiv:2410.04514*, 2024. 1

[12] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 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*, pages 13418–13427, 2024. 7

[13] Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27036–27046, 2024. 2

[14] Nick Jiang, Anish Kachinthaya, Suzie Petryk, and Yossi Gandelsman. Interpreting and editing vision-language representations to mitigate hallucinations. *arXiv preprint arXiv:2410.02762*, 2024. 7

[15] Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25004–25014, 2025. 1, 3, 7

[16] Omri Kaduri, Shai Bagon, and Tali Dekel. What's in the image? a deep-dive into the vision of vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14549–14558, 2025. 1

[17] Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. *arXiv preprint arXiv:2503.03321*, 2025. 1, 3

[18] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882, 2024. 1, 3, 7

[19] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1, 2

[20] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 1

[21] Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori B Hashimoto, Luke Zettlemoyer, and

Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312, 2023. 3

[22] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 1

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 3

[24] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023. 2

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1, 2

[26] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024. 2, 3, 8

[27] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024. 1

[28] Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, pages 125–140. Springer, 2024. 1, 3, 5, 7, 8

[29] Xinyu Lyu, Beitao Chen, Lianli Gao, Hengtao Shen, and Jingkuan Song. Alleviating hallucinations in large vision-language models through hallucination-induced optimization. *Advances in Neural Information Processing Systems*, 37:122811–122832, 2024. 2

[30] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 6

[31] nostalgebraist. Interpreting gpt: The logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens, 2020. 2, 4

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[33] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018. 4, 6, 7

[34] Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Ö Arık, and Tomas Pfister. Mitigating object hallucination in mllms via data-augmented phrase-level alignment. *arXiv preprint arXiv:2405.18654*, 2024. 2

[35] Zailong Tian, Zhuoheng Han, Yanzhe Chen, Haozhe Xu, Xi Yang, Richeng Xuan, Houfeng Wang, and Lizi Liao. Overconfidence in llm-as-a-judge: Diagnosis and confidence-driven solution. *arXiv preprint arXiv:2508.06225*, 2025. 7

[36] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578, 2024. 1, 3

[37] Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. Mllm can see? dynamic correction decoding for hallucination mitigation. *arXiv preprint arXiv:2410.11779*, 2024. 3, 7

[38] Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, et al. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023. 6, 7

[39] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 8

[40] Huyu Wu, Meng Tang, Xinhan Zheng, and Haiyun Jiang. When language overrules: Revealing text dominance in multimodal large language models. *arXiv preprint arXiv:2508.10552*, 2025. 1, 3

[41] Zhihe Yang, Xufang Luo, Dongqi Han, Yunjian Xu, and Dongsheng Li. 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*, pages 10610–10620, 2025. 2

[42] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 13040–13051, 2024. 2

[43] Hao Yin, Guangzong Si, and Zilei Wang. Clearsight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14625–14634, 2025. 1, 3

[44] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 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*, pages 13807–13816, 2024. 2

[45] Linxi Zhao, Yihe Deng, Weitong Zhang, and Quanquan Gu. Mitigating object hallucination in large vision-language models via classifier-free guidance. *arXiv e-prints*, pages arXiv–2402, 2024. 2, 3

[46] Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhanc-

ing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*, 2023. 2

[47] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*, 2023. 7