WHO IS THE REAL WITNESS? LOCALIZING ATTENTION HEADS THROUGH IMAGE-INDUCED SEMANTIC SHIFTS IN LVLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large vision-language models (LVLMs) have demonstrated strong performance on multimodal tasks. However, they may overrely on prior knowledge when it conflicts with visual information, producing plausible predictions that are inconsistent with the input image. Previous studies have sought to mitigate this issue by identifying and enhancing attention heads that are sensitive to visual input. We observe that visually sensitive heads do not always capture the key cues needed for the current prediction. In this paper, we propose **W**itness **I**dentification **T**hrough **N**ext-token **E**vidence of **S**emantic **S**hifts (WITNESS), a training-free method. It projects each head's outputs into the vocabulary space and compares the resulting predictive distributions under text-only and image-text conditions over the same candidate token set. The resulting image-induced semantic shifts are quantified using Jensen–Shannon (JS) divergence. WITNESS then selects and enhances heads exhibiting larger shifts and combines this intervention with contrastive decoding. We conduct extensive experiments on the counterfactual datasets WHOOPS-AHA! and SynConFact using four LVLMs. Compared with the base models, WITNESS achieves relative improvements of 6.49% and 16.16% in accuracy averaged across models on the two datasets, respectively. Our code is available at https://anonymous.4open.science/r/WITNESS-57B7.

1 INTRODUCTION

Large vision-language models (LVLMs) (Liu et al., 2024a; Wang et al., 2024; Bai et al., 2025; Chen et al., 2025) have demonstrated strong performance on multimodal tasks such as visual question answering (Cocchi et al., 2025) and image understanding (Tian et al., 2026). These models combine prior knowledge acquired during pretraining with visual information from the input image (Wu et al., 2025c; Zhou et al., 2025). However, when these sources of information conflict, overreliance on prior knowledge may cause models to overlook visual information and produce plausible predictions that are inconsistent with the image (He et al., 2025; Ortu et al., 2026). Such inconsistencies can undermine the accuracy and reliability of LVLMs across vision-language tasks, limiting their practical applications (Sun et al., 2026a; Yin et al., 2026).

To mitigate this issue, existing approaches can be broadly categorized into two groups: training-based and training-free methods. Training-based methods improve the faithfulness of model outputs to visual information by refining multimodal representations (Agrawal et al., 2026) or applying preference optimization (Liu et al., 2025a; Wu et al., 2025a; Lim et al., 2026). These methods require additional parameter optimization and incur corresponding training costs. Training-free methods improve visual faithfulness through inference-time interventions. Some methods contrast predictive distributions under different conditions to reduce models' overreliance on prior knowledge (Deng & Yang, 2025; Ren et al., 2026; Bi et al., 2026). However, directly suppressing prior knowledge may compromise the coherence and accuracy of generated content (Yin et al., 2025). Other methods identify attention heads sensitive to visual input and intervene on these heads during generation (He et al., 2025). HAVAE further excludes the influence of inert visual tokens when identifying attention heads for enhancement (Chen et al., 2026b).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: Resolving conflicts between visual information and prior knowledge. In the golf example on the left and the chess example on the right, the Base LVLM overrelies on prior knowledge, causing it to overlook visual information and produce predictions inconsistent with the image. WITNESS enhances attention heads that focus on the key visual cues needed for the current prediction, resulting in predictions that are more consistent with the visual information in the image.

In this work, we observe that attention heads sensitive to visual input attend to different image regions. As shown in Fig. 1, Head 20 in the golf example and Head 5 in the chess example attend to the key visual cues needed for the current prediction, while Head 16 and Head 23 attend to other visual information with limited relevance to the current prediction. We further find that enhancing heads with limited relevance to the current prediction may interfere with the model's decisions, as analyzed in Section 3.1. Based on these observations, we propose WITNESS, a training-free method that localizes and enhances attention heads by measuring image-induced changes in their preferences over the current candidate token set, as shown in Fig. 2. Specifically, we compare each head's predictive distributions under text-only and image-text conditions over the same candidate token set. We refer to the resulting change as an image-induced semantic shift and quantify it using Jensen–Shannon (JS) divergence. We then select and enhance heads exhibiting larger semantic shifts and combine this intervention with contrastive decoding to encourage the model to make more effective use of the key visual cues needed for the current prediction. The main contributions of our paper are:

- We find that sensitivity to visual input does not necessarily indicate that an attention head captures the key visual cues needed for the current prediction. Our analysis further shows that enhancing heads with limited relevance to the current prediction may interfere with the model's decisions.
- We propose WITNESS, a training-free method that localizes and enhances attention heads based on image-induced semantic shifts over the same candidate token set. WITNESS combines this intervention with contrastive decoding to better utilize visual information for the current prediction.
- We construct SynConFact, a vision-language dataset for studying conflicts between visual information and prior knowledge through next-token prediction. It pairs generated counterfactual images with textual prompts and factual and counterfactual candidate words.
- We conduct extensive experiments on two counterfactual datasets using four LVLMs. Compared with the base models, WITNESS achieves relative improvements of 6.49% and 16.16% in accuracy averaged across these models on WHOOPS-AHA! and SynConFact, respectively.

2 Related Work

2.1 Large Vision-Language Models

Large vision-language models (LVLMs) have advanced rapidly in recent years, demonstrating strong performance on tasks such as visual question answering and image understanding (Caffagni et al., 2024; Fu et al., 2025b; Li et al., 2025; Sun et al., 2026b). These models typically combine visual encoders with pretrained language models (Li et al., 2023; Liu et al., 2023; Zhu et al., 2024; Bi et al., 2025), enabling them to integrate visual information from the input image with prior knowledge acquired during pretraining. LLaVA-NeXT enhances visual perception through higher input

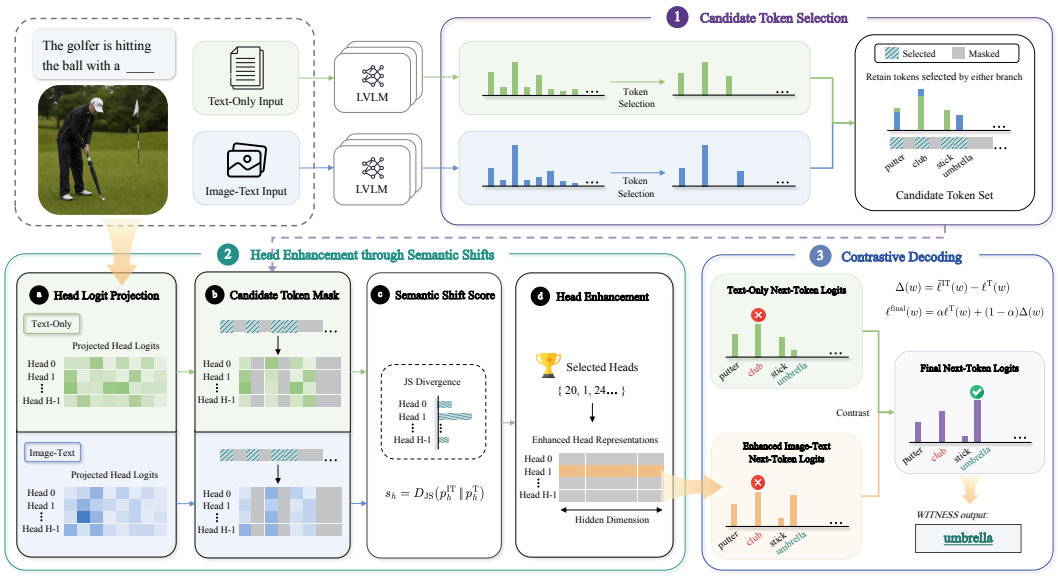


Figure 2: Overview of the WITNESS framework.

image resolution and an improved visual instruction tuning data mixture (Liu et al., 2024a). Qwen2-VL introduces dynamic resolution processing to represent images with variable numbers of visual tokens (Wang et al., 2024), while Qwen3-VL integrates multi-level visual features through DeepStack to strengthen vision-language alignment (Bai et al., 2025). Janus-Pro improves multimodal understanding and generation through optimized training strategies and increased data and model scales (Chen et al., 2025). However, when visual information conflicts with prior knowledge, models may overrely on prior knowledge and consequently overlook the visual information in the current image (Liu et al., 2024b; He et al., 2025).

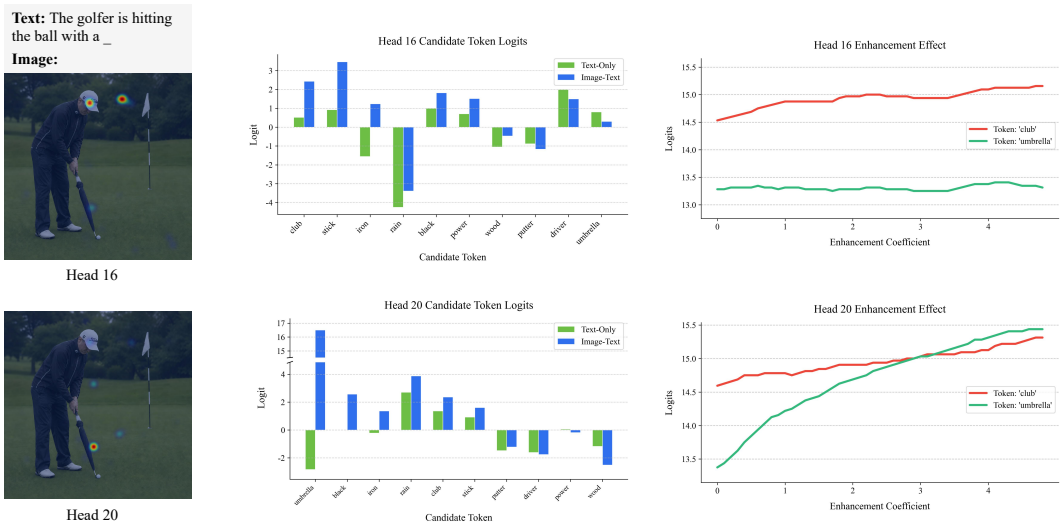
2.2 Knowledge Conflicts in LVLMs

Various approaches have been proposed to improve LVLMs' use of visual information and mitigate their overreliance on prior knowledge (Wu et al., 2025b; Li et al., 2026). These approaches can be broadly categorized into two groups: training-based and training-free methods. Training-based methods improve models' use of visual information through multimodal representation refinement and alignment (Agrawal et al., 2026; Chen et al., 2026a) or preference optimization (Liu et al., 2025a; Wu et al., 2025a; Lim et al., 2026). These methods require additional parameter updates. Training-free methods intervene during inference (Sarkar et al., 2025; Fazli et al., 2026). Contrastive decoding calibrates predictions by comparing outputs under different conditions (Leng et al., 2024; Park et al., 2025; Wang et al., 2026). In textual knowledge conflicts, CK-PLUG (Bi et al., 2026) uses entropy changes to regulate reliance on parametric and contextual knowledge. Other methods directly intervene on attention heads (Tang et al., 2026; Chen et al., 2026b; Jiang et al., 2025; Ji et al., 2026). VHR (He et al., 2025) identifies and reinforces vision-aware heads based on differences in head outputs with and without visual input. Ortu et al. (2026) compare factual and counterfactual token logits to select heads for steering knowledge reliance. In contrast, WITNESS selects and enhances attention heads based on image-induced changes in their token preferences within the current candidate token set. This selective enhancement aims to improve the use of key visual cues needed for the current prediction.

3 Preliminary Study

3.1 Image-Induced Semantic Shifts

In this section, we examine the roles of attention heads sensitive to visual input in the current prediction. As shown in Fig. 3(a), Head 16 primarily attends to the region around the golfer's hat, which



(a) Attention maps **(b)** Image-induced semantic shifts **(c)** Head enhancement effect

Figure 3: Image-induced semantic shifts and head enhancement effects in Qwen2-VL.

has limited relevance to the current prediction. In contrast, Head 20 focuses on the umbrella, the key visual cue needed for this prediction. Appendix A.2 provides attention maps for additional heads.

Fig. 3(b) presents the candidate token logits obtained through Head Logit Projection under text-only and image-text conditions. Appendix A.3 provides evidence supporting the comparison of heads with the same index across the two conditions. With the image introduced, the candidate token logits of Head 16 show only limited changes. Tokens consistent with prior knowledge, such as *club*, remain relatively high. In contrast, the logit of *umbrella* increases substantially in Head 20, leading to a clear change in its candidate token preferences toward the image content. We refer to these changes in candidate token preferences as image-induced semantic shifts. Appendix A.4 presents these shifts over the complete candidate token set.

To further examine how these heads affect the current prediction, we enhance each head separately, as shown in Fig. 3(c). Stronger enhancement of Head 16 generally raises the logit for *club*, with little change for *umbrella*. Enhancing Head 20 instead raises the logit for *umbrella*, which exceeds that of *club* at larger coefficients. These observations motivate using image-induced semantic shifts within the candidate token set to identify heads relevant to the current prediction.

3.2 JS DIVERGENCE QUANTIFICATION

We use JS divergence (Lin, 1991; Zhou et al., 2025) to quantify the image-induced semantic shifts observed in the previous section. As shown in Fig. 4, the JS divergences of Head 16 and Head 20 over the full vocabulary are 0.6927 and 0.6921, respectively, making them nearly indistinguishable. However, restricting the computation to the current candidate token set reduces Head 16's JS divergence to 0.2201, while Head 20 retains a high divergence of 0.6885, revealing a clearer difference between the two heads. These results indicate that a large image-induced semantic shift over the full vocabulary does not necessarily imply that a head captures the key visual cues needed for the current prediction. Focusing the comparison on the current candidate token set helps identify attention heads relevant to the current prediction, providing empirical support for head selection based on semantic shifts within this set.

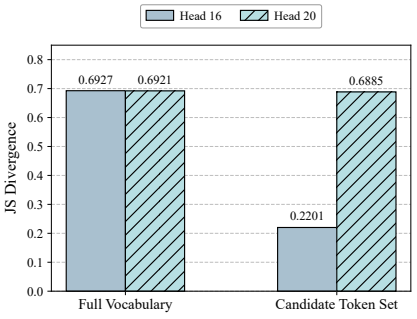


Figure 4: Effect of candidate token set restriction on head-wise JS divergence in Qwen2-VL.

4 METHOD

In this section, we introduce WITNESS, a training-free method that localizes and enhances attention heads in LVLMs based on image-induced semantic shifts. As shown in Fig. 2, WITNESS consists of three components. First, we construct a candidate token set from tokens selected with and without the image. Next, we use head-wise semantic shift scoring to identify the heads most responsive to key visual cues and enhance them. Finally, we apply contrastive decoding to obtain the final logits.

4.1 CANDIDATE TOKEN SELECTION

We consider next-token prediction given an image I and a textual prompt T. As discussed in Section 3.2, semantic shift scores computed over the full vocabulary may not sufficiently distinguish preference changes across attention heads. We therefore construct a shared candidate token set for continuing the current text from the model's next-token predictions with and without visual input.

Let $\boldsymbol{\ell}^c \in \mathbb{R}^{|\mathcal{V}|}$ denote the next-token logits produced by the same LVLM over vocabulary $\mathcal{V}$ under condition $c \in \{\mathrm{IT}, \mathrm{T}\}$, where IT and T denote the image-text and text-only conditions, respectively. For each condition c, we define a selection threshold as

$$\theta_c = \min\left(\ell^c_{(k)}, \max_{u\in\mathcal{V}} \ell^c(u) + \log\rho\right), \tag{1}$$

where $\ell^c(u)$ denotes the logit of token u, $\ell^c_{(k)}$ is the k-th largest logit, and $\rho \in (0, 1]$ controls the relative probability threshold. We then construct the candidate token set by taking the union of the tokens selected under the two conditions:

$$\mathcal{C}_c = \{w \in \mathcal{V} : \ell^c(w) \geq \theta_c\}, \qquad \mathcal{C} = \mathcal{C}_{\mathrm{IT}} \cup \mathcal{C}_{\mathrm{T}}. \tag{2}$$

Taking the union retains tokens that satisfy the selection rule under either input condition. The resulting set is used in the Candidate Token Mask step of the next subsection to restrict semantic shift scoring to these candidates.

4.2 HEAD ENHANCEMENT THROUGH SEMANTIC SHIFTS

To identify attention heads exhibiting substantial image-induced semantic shifts within the current candidate token set, we first project individual head outputs into the vocabulary space. We then compare their candidate prediction distributions across the two input conditions, using the resulting semantic shift scores to select and enhance heads.

Head Logit Projection. We extract the attention head outputs at the final input position in the last language model layer, before the attention output projection. Let d denote the model's hidden dimension, H the number of attention heads in this layer, and $d_{\mathrm{head}} = d/H$ the dimension of each head. For input condition $c \in \{\mathrm{IT}, \mathrm{T}\}$, let $\mathbf{x}^c_h \in \mathbb{R}^{d_{\mathrm{head}}}$ denote the output of attention head h.

Let $\mathbf{W}^O_h \in \mathbb{R}^{d\times d_{\mathrm{head}}}$ denote the column block of the attention output projection matrix $\mathbf{W}^O$ corresponding to head h. We first map the head output into the model's hidden space:

$$\mathbf{z}^c_h = \mathbf{W}^O_h \mathbf{x}^c_h \in \mathbb{R}^d. \tag{3}$$

We then apply the model's final normalization layer and vocabulary output layer to obtain the projected head logits:

$$\boldsymbol{\ell}^c_h = \mathrm{LMHead}\left(\mathrm{Norm}\left(\mathbf{z}^c_h\right)\right) \in \mathbb{R}^{|\mathcal{V}|}. \tag{4}$$

Candidate Token Mask. Using the candidate token set $\mathcal{C}$ defined in Equation 2, we apply the same mask to the projected logits $\boldsymbol{\ell}^c_h$ of all attention heads under both input conditions. Tokens outside $\mathcal{C}$ are masked, and the remaining logits are normalized using softmax:

$$p^c_h(w) = \begin{cases} \dfrac{\exp(\ell^c_h(w))}{\sum_{u\in\mathcal{C}} \exp(\ell^c_h(u))}, & w \in \mathcal{C}, \\ 0, & \text{otherwise}, \end{cases} \qquad c \in \{\mathrm{IT}, \mathrm{T}\}. \tag{5}$$

Semantic Shift Score. We use Jensen–Shannon (JS) divergence to quantify the difference between the candidate prediction distributions of the same attention head with and without visual input. We first define their average distribution:

$$m_h = \frac{1}{2}\left(p_h^{\mathrm{IT}} + p_h^{\mathrm{T}}\right). \tag{6}$$

The semantic shift score of attention head h is then defined as:

$$s_h = \frac{1}{2} D_{\mathrm{KL}}\left(p_h^{\mathrm{IT}} \,\|\, m_h\right) + \frac{1}{2} D_{\mathrm{KL}}\left(p_h^{\mathrm{T}} \,\|\, m_h\right), \tag{7}$$

where D_{KL} denotes Kullback–Leibler divergence. A larger score indicates a greater change in the head's prediction distribution within the candidate token set after introducing visual input. We select heads for enhancement according to the ratio of each head's score to the maximum score for the current input:

$$\mathcal{H}_{\mathrm{sel}} = \left\{ h : \frac{s_h}{\max_j s_j} \geq \tau \right\}. \tag{8}$$

Here, j ranges over all attention heads in this layer, and τ is the selection threshold. We denote the normalized ratio above by $\widehat{s}_h$ and use it for subsequent head enhancement.

Head Enhancement. Under the image-text condition, we rescale the selected head outputs at the final input position in the last language model layer using their normalized semantic shift scores:

$$\widetilde{\mathbf{x}}_h^{\mathrm{IT}} = \begin{cases} (1 + \lambda \widehat{s}_h)\, \mathbf{x}_h^{\mathrm{IT}}, & h \in \mathcal{H}_{\mathrm{sel}}, \\ \mathbf{x}_h^{\mathrm{IT}}, & \text{otherwise}, \end{cases} \tag{9}$$

where λ controls the enhancement strength. The modified head outputs are propagated through the remaining model computation to obtain the enhanced image-text next-token logits $\widetilde{\boldsymbol{\ell}}^{\mathrm{IT}}$.

4.3 Contrastive Decoding

To improve the faithfulness of model outputs to visual information when it conflicts with prior knowledge, we apply contrastive decoding using the enhanced image-text logits and the text-only logits.

For each token in the candidate token set $\mathcal{C}$ defined in Equation 2, we define a contrastive signal as the difference between the enhanced image-text logits and the text-only logits:

$$\Delta(w) = \widetilde{\ell}^{\mathrm{IT}}(w) - \ell^{\mathrm{T}}(w), \qquad w \in \mathcal{C}. \tag{10}$$

We combine the text-only logits with the contrastive signal to obtain the final next-token logits:

$$\ell^{\mathrm{final}}(w) = \begin{cases} \alpha \ell^{\mathrm{T}}(w) + (1-\alpha)\Delta(w), & w \in \mathcal{C}, \\ -\infty, & \text{otherwise}, \end{cases} \tag{11}$$

where $\alpha \in [0, 1]$ controls the relative contributions of the text-only logits and the contrastive signal. By combining head enhancement with contrastive decoding, we aim to help the model make better use of the key visual cues needed for the current prediction.

5 Experimental Setup

5.1 Datasets and Evaluation Metrics

Datasets. We primarily evaluate WITNESS on WHOOPS-AHA! (Ortu et al., 2026) and our synthetic counterfactual dataset SynConFact, to study conflicts between visual information and prior knowledge. WHOOPS-AHA! contains 500 instances built from images in WHOOPS! (Bitton-Guetta et al., 2023) that defy commonsense knowledge. SynConFact contains 420 instances evenly distributed across six categories: Animal, Attribute, Historical, Material, Object, and Scene. Each instance in both datasets consists of a counterfactual image, an incomplete textual prompt, and annotated factual and counterfactual reference answers. The factual answers reflect prior knowledge,

Table 1: Performance comparison on WHOOPS-AHA! and SynConFact.

Model	Method	WHOOPS -AHA!	SynConFact						
			Animal	Attribute	Historical	Material	Object	Scene	Average
Janus-Pro	Baseline	62.2	91.4	80.0	65.7	64.3	58.6	42.9	67.1
	VHR	61.6	90.0	78.6	65.7	62.9	60.0	40.0	66.2
	CK-PLUG	68.2	92.9	88.6	74.3	74.3	75.7	65.7	78.6
	Ours	**69.0**	92.9	**90.0**	**75.7**	**77.1**	**78.6**	**67.1**	**80.2**
LLaVA-NeXT	Baseline	65.4	92.9	74.3	68.6	67.1	71.4	32.9	67.9
	VHR	65.0	91.4	74.3	67.1	65.7	72.9	35.7	67.9
	CK-PLUG	71.6	98.6	75.7	81.4	78.6	88.6	52.9	79.3
	Ours	**72.2**	98.6	**78.6**	**82.9**	**80.0**	88.6	**54.3**	**80.5**
Qwen2-VL	Baseline	73.4	94.3	75.7	82.9	84.3	85.7	24.3	74.5
	VHR	72.6	91.4	74.3	82.9	81.4	85.7	40.0	76.0
	CK-PLUG	75.2	**95.7**	88.6	88.6	94.3	91.4	57.1	86.0
	Ours	**75.8**	94.3	**90.0**	88.6	**97.1**	**94.3**	**68.6**	**88.8**
Qwen3-VL	Baseline	76.2	95.7	77.1	84.3	87.1	92.9	47.1	80.7
	VHR	75.2	94.3	78.6	84.3	85.7	91.4	44.3	79.8
	CK-PLUG	77.6	95.7	81.4	88.6	90.0	92.9	70.0	86.4
	Ours	**78.2**	**97.1**	81.4	88.6	**94.3**	92.9	**71.4**	**87.6**

while the counterfactual answers are consistent with the visual information. The detailed construction procedure for SynConFact is provided in Appendix A.1. Additionally, we evaluate Janus-Pro and Qwen2-VL on MME (Fu et al., 2025a) to assess the effect of WITNESS on general perception and cognition capabilities.

Evaluation Metrics. For WHOOPS-AHA! (Ortu et al., 2026) and SynConFact, we report accuracy (%). To accommodate variations in response format, we consider a response correct if any annotated counterfactual reference answer appears anywhere in the full generated response. For MME (Fu et al., 2025a), we follow the official evaluation protocol and compute each subtask score as the sum of accuracy and accuracy+, consistent with prior work (Woo et al., 2024; Liu et al., 2025b). We report the perception and cognition scores, together with their sum as the overall score.

5.2 Baselines

We evaluate WITNESS on four representative LVLMs: Janus-Pro-7B (Chen et al., 2025), LLaVA-NeXT-7B (Liu et al., 2024a), Qwen2-VL-7B (Wang et al., 2024), and Qwen3-VL-8B (Bai et al., 2025). We use the base models with greedy decoding as the baseline. We additionally compare WITNESS with two inference-time intervention methods: VHR (He et al., 2025) and CK-PLUG (Bi et al., 2026). VHR identifies and reinforces attention heads based on their sensitivity to visual input. CK-PLUG controls the model's reliance on parametric and contextual knowledge by adjusting token prediction distributions. Additional experimental settings are provided in Appendix A.5.

5.3 Hyperparameter Settings

We apply attention head interventions only in the last language model layer. For candidate token set construction, we set the relative probability threshold to $\rho = 0.1$ and retain at least $k = 10$ tokens under each input condition. We set $\alpha = 0.4$ for contrastive decoding. We select the head selection threshold τ and enhancement coefficient λ through parameter sweeps on WHOOPS-AHA!, as reported in Section 6.3. The resulting settings are $\tau = 0.8$ for all models and $\lambda = 1$ for Janus-Pro, $\lambda = 2$ for LLaVA-NeXT and Qwen2-VL, and $\lambda = 3$ for Qwen3-VL. For each model, we apply the same hyperparameter settings to SynConFact and MME without further tuning.

6 EXPERIMENTS

6.1 MAIN RESULTS

Table 1 shows that WITNESS achieves the highest accuracy on WHOOPS-AHA! and the highest average accuracy on SynConFact across all four LVLMs. Compared with the base models, WITNESS achieves relative improvements of 6.49% and 16.16% in accuracy averaged across models on the two datasets, respectively. VHR does not consistently outperform the base models under these evaluation settings. CK-PLUG improves performance on both datasets, while WITNESS further improves upon CK-PLUG for every evaluated model. The gains over CK-PLUG range from 0.6 to 0.8 percentage points on WHOOPS-AHA! and from 1.2 to 2.8 percentage points on SynConFact. These results suggest that when visual information conflicts with prior knowledge, WITNESS helps models make more effective use of the visual information needed for the current prediction.

The improvements achieved by WITNESS vary across SynConFact categories. WITNESS performs particularly well in the Material and Scene categories, outperforming all competing methods in both categories across all four LVLMs. The gains over the base models in the Animal category are relatively modest, possibly because these tasks involve relatively simple knowledge and the base models already achieve high accuracy. Compared with the base models, the Scene category shows the largest accuracy gain for each of the four LVLMs. For Qwen2-VL, accuracy in this category increases from 24.3% to 68.6%. This also exceeds the accuracy achieved by CK-PLUG by 11.5 percentage points. This result indicates that WITNESS reduces reliance on common scene associations by leveraging key visual cues.

6.2 RESULTS ON MME

Table 2: Evaluation results on the MME benchmark.

Model	Method	Perception	Cognition	Overall
Janus-Pro	Baseline	1501.1	281.1	1782.2
	Ours	**1520.1**	**282.1**	**1802.3**
Qwen2-VL	Baseline	**1694.4**	**621.8**	**2316.2**
	Ours	1684.2	606.8	2291.0

We evaluate Janus-Pro and Qwen2-VL on MME to examine the effect of WITNESS on general multimodal performance. As shown in Table 2, WITNESS increases both perception and cognition scores for Janus-Pro. However, both scores decrease slightly for Qwen2-VL. More detailed results are provided in Table 7 in Appendix A.6. Janus-Pro improves on subtasks such as color and scene recognition, but shows small declines in position recognition and commonsense reasoning. Qwen2-VL also improves on scene, landmark, and artwork recognition, despite declines in subtasks such as OCR and numerical calculation. These results suggest that the effect of WITNESS on general multimodal capabilities varies across models, while the relative changes in overall MME scores remain modest.

6.3 FURTHER ANALYSIS

Head Selection Strategies. We compare our head selection strategy with Random and Fixed Top-2 on WHOOPS-AHA! using Janus-Pro and Qwen2-VL. Random selects attention heads at random, while Fixed Top-2 selects the two heads with the highest semantic shift scores for each input. Our method selects heads whose normalized scores meet the relative threshold, allowing the number of selected heads to vary across inputs. As shown in Table 3, the threshold rule used in WITNESS achieves the highest accuracy among the three strategies on both models. These results support using semantic shift scores to guide head selection.

Sensitivity to the Head Selection Threshold. We examine the sensitivity of WITNESS to the relative head selection threshold τ on WHOOPS-AHA! using Janus-Pro and Qwen2-VL. We vary τ over $\{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$. A larger threshold makes head selection more restrictive; at $\tau = 1.0$, only heads attaining the maximum semantic shift score are retained. As shown in Table 4, both models achieve their highest accuracy at $\tau = 0.8$. These results suggest that retaining multiple heads with substantial semantic shifts may improve performance. However, further relaxing the selection criterion may also include heads that offer limited benefit for the current prediction and reduce the performance gains achieved by the method.

Table 3: Comparison of head selection strategies on WHOOPS-AHA!. Avg. #Heads denotes the average number of selected heads per input.

Method	Janus-Pro		Qwen2-VL	
	Acc.	Avg. #Heads	Acc.	Avg. #Heads
Random	68.2	2.05	74.6	2.36
Fixed Top-2	68.6	2.00	75.2	2.00
Ours	**69.0**	2.05	**75.8**	2.36

Table 4: Sensitivity to the relative head selection threshold τ on WHOOPS-AHA!. Columns indicate different values of the selection threshold, and results are reported as accuracy (%) for Janus-Pro and Qwen2-VL.

Model	0.0	0.2	0.4	0.6	0.8	1.0
Janus-Pro	68.2	68.2	68.8	68.6	**69.0**	68.6
Qwen2-VL	75.2	75.4	75.6	75.6	**75.8**	75.6

Sensitivity to the Enhancement Coefficient. We examine the sensitivity of WITNESS to the enhancement coefficient λ on WHOOPS-AHA! using four LVLMs. We vary λ over $\{-1, 0, 1, 2, 3, 4, 5, 6\}$. At $\lambda = 0$, the attention head outputs remain unchanged; at $\lambda = -1$, the selected head outputs are suppressed. As shown in Fig. 5, moderate positive coefficients improve accuracy compared with $\lambda = 0$, while accuracy generally declines as the coefficient increases further. Among the tested values, the coefficient yielding the highest accuracy differs across models. These results indicate that different models respond differently to attention head enhancement, supporting using a separate enhancement coefficient for each model rather than a single shared value.

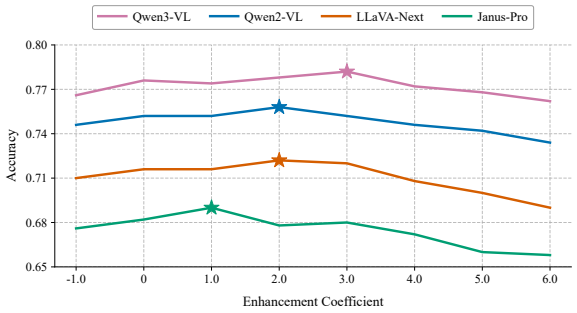


Figure 5: Sensitivity to the enhancement coefficient.

Ablation Studies. We conduct ablation studies on WHOOPS-AHA! to examine the contributions of contrastive decoding and the candidate token mask. As shown in Table 5, removing contrastive decoding reduces accuracy by an average of 3.95 percentage points across models relative to the full method. Removing the candidate token mask also decreases accuracy for all evaluated models, with an average reduction of 1.10 percentage points. These results indicate that both components help improve the performance of WITNESS.

Table 5: Ablation study of the proposed components.

Method	Janus-Pro	LLaVA-NeXT	Qwen2-VL	Qwen3-VL
Baseline	62.2	65.4	73.4	76.2
w/o Contrastive Decoding	62.6	66.2	73.6	77.0
w/o Candidate Token Mask	68.2	71.2	74.4	77.0
Full Method	**69.0**	**72.2**	**75.8**	**78.2**

7 Conclusion

In this paper, we presented WITNESS, a training-free method that localizes and enhances attention heads in LVLMs based on image-induced semantic shifts. This method helps mitigate conflicts between visual information and prior knowledge in LVLMs. We also constructed SynConFact, a dataset for studying visual counterfactuals through next-token prediction. Our experiments show that, compared with the base models, WITNESS achieves relative improvements of 6.49% and 16.16% in accuracy averaged across the evaluated LVLMs on WHOOPS-AHA! and SynConFact, respectively. Our results support considering each head's relevance to the visual information needed for the current prediction during enhancement, rather than relying solely on its sensitivity to visual input. These findings provide an empirical basis for further investigation into how different attention heads contribute to the use of visual information in LVLMs.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic data sets, designing or providing feedback on research methodology or experiments, implementing methods, assisting with translation, and interpreting results. We have not used generative AI tools for developing theoretical models or conceptual frameworks, formulating mathematical claims, proposing or refining hypotheses, cleaning and reformatting datasets, or supporting qualitative and thematic data analysis, and providing critical ingredients for proving mathematical claims and assisting in the writing of proofs are not applicable to this work. Additionally, we used generative AI tools for summarizing or analyzing existing literature, sourcing and searching for information, drafting parts of the research paper, editing the research paper to improve readability, and identifying relevant literature. We have reviewed all AI-assisted work. AI-generated synthetic data were manually inspected by the authors before being used in our experiments. Suggestions related to research methodology and experimental design were treated as advisory and were adopted only after evaluation by the authors. AI-assisted code and method implementations were manually reviewed, executed, and tested for correctness. AI-assisted translations were checked against the original text, and AI-assisted interpretations of experimental results were verified against the actual experimental outputs and quantitative evidence. For literature-related assistance, including literature summarization, information searching, and identification of relevant work, we verified the information against the original papers and other primary sources. All AI-assisted drafts and edits to the manuscript were manually reviewed and revised by the authors to ensure accuracy, clarity, and consistency with the underlying research. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This research was conducted in accordance with established ethical standards for scientific inquiry. We considered issues related to dataset usage and release, potential societal harms, fairness and bias, privacy and security, legal compliance, and research integrity. Our study does not involve the recruitment or interaction of human participants or the collection of private personally identifiable information, and therefore no Institutional Review Board (IRB) approval was required. The experiments are conducted using existing benchmark datasets and our publicly released SynConFact dataset, which consists of synthetically generated counterfactual images and manually curated textual annotations for research purposes. Potential risks related to model bias, fairness, privacy, and unintended misuse were considered when constructing the dataset and designing and interpreting the experiments. We publicly release SynConFact to support transparency and reproducibility, and strive to ensure accuracy, transparency, and reproducibility in the presentation of our datasets, methodology, experimental results, and conclusions.

REPRODUCIBILITY STATEMENT

We have made several efforts to support the reproducibility of this work. The main paper provides a detailed description of the proposed method, experimental settings, evaluation protocols, and benchmark results. The appendix provides additional implementation details, dataset construction and processing procedures, and analysis results. An anonymized version of the SynConFact dataset, together with the source code and evaluation scripts required to reproduce the experiments, is provided in the supplementary materials. These resources are intended to facilitate reproduction of the dataset preparation, head selection and enhancement procedures, and the experimental results reported in this work.

REFERENCES

Aakriti Agrawal, Gouthaman KV, Rohith Aralikatti, Gauri Jagatap, Jiaxin Yuan, Sarvesh Baskar, Vijay Kamarshi, Andrea Fanelli, and Furong Huang. Towards mitigating hallucinations in large vision-language models by refining textual embeddings. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 42032–42048, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.2086. URL https://aclanthology.org/2026.findings-acl.2086/.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Baolong Bi, Shenghua Liu, Yiwei Wang, Yilong Xu, Junfeng Fang, Lingrui Mei, and Xueqi Cheng. Parameters vs. context: Fine-grained control of knowledge reliance in language models. In *International Conference on Learning Representations*, volume 2026, pp. 106786–106808, 2026.

Jing Bi, Junjia Guo, Yunlong Tang, Lianggong Bruce Wen, Zhang Liu, Bingjie Wang, and Chenliang Xu. Unveiling visual perception in language models: An attention head analysis approach. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4135–4144. IEEE, 2025.

Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2616–2627. IEEE, 2023.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The revolution of multimodal large language models: A survey. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13590–13618, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.807. URL https://aclanthology.org/2024.findings-acl.807/.

Linqing Chen, Hanmeng Zhong, Wentao Wu, and Peng Zhou. The dominance of text space: Unveiling the asymmetric nature of cross-modal alignment in large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 36666–36678, 2026a.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

Yangneng Chen, Junlin Li, Weijun Yao, Xilai Ma, Guodong Du, Wenya Wang, and Jing Li. Vocabulary hijacking in lvlms: Unveiling critical attention heads by excluding inert tokens to mitigate hallucination. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 38458–38485, 2026b.

Federico Cocchi, Nicholas Moratelli, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Augmenting multimodal llms with self-reflective tokens for knowledge-based visual question answering. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9199–9209. IEEE, 2025.

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

Mehrdad Fazli, Bowen Wei, and Ziwei Zhu. Inject to heal: Alleviating hallucination in lvlms via context embedding injection. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 41177–41193, 2026.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38, 2025a.

Xingyu Fu, Minqian Liu, Zhengyuan Yang, John Corring, Yijuan Lu, Jianwei Yang, Dan Roth, Dinei Florencio, and Cha Zhang. Refocus: Visual editing as a chain of thought for structured image understanding. *arXiv preprint arXiv:2501.05452*, 2025b.

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

Junyang Ji, Qifan Liu, Wenming Yang, and Zhihai He. Causallens: Sensitivity-guided multi-head causal intervention for hallucination mitigation in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4199–4209, 2026.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25004–25014. IEEE, 2025.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882. IEEE, 2024.

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*, pp. 19730–19742. PmLR, 2023.

Xinwei Li, Li Lin, Hui Jiao, Li Yao, Tien-Tsin Wong, and Hanqian Wu. Dive: Decoupling intra-layer visual evidence for mitigating hallucinations in large vision-language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 37552–37568, 2026.

Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1578–1597. IEEE, 2025.

Byeonggeuk Lim, JungMin Yun, Junehyoung Kwon, Kyeonghyun Kim, and YoungBin Kim. Aligning with your own voice: Self-corrected preference learning for hallucination mitigation in lvlms. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 35829–35849, 2026.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

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

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024a.

Wenqi Liu, Xuemeng Song, Jiaxi Li, Yinwei Wei, Na Zheng, Jianhua Yin, and Liqiang Nie. Mitigating hallucination through theory-consistent symmetric multimodal preference optimization. *Advances in Neural Information Processing Systems*, 38:111259–111284, 2025a.

Xiaoyuan Liu, Wenxuan Wang, Youliang Yuan, Jen-tse Huang, Qiuzhi Liu, Pinjia He, and Zhaopeng Tu. Insight over sight: Exploring the vision-knowledge conflicts in multimodal llms. *arXiv preprint arXiv:2410.08145*, 2024b.

Xixi Liu, Ailin Deng, and Christopher Zach. Energy-guided decoding for object hallucination mitigation. *arXiv preprint arXiv:2507.07731*, 2025b.

Francesco Ortu, Zhijing Jin, Diego Doimo, and Alberto Cazzaniga. When seeing overrides knowing: Disentangling knowledge conflicts in vision-language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14109–14130, 2026.

Woohyeon Park, Woojin Kim, Jaeik Kim, and Jaeyoung Do. Second: Mitigating perceptual hallucination in vision-language models via selective and contrastive decoding. *arXiv preprint arXiv:2506.08391*, 2025.

Lingfeng Ren, Weihao Yu, Runpeng Yu, and Xinchao Wang. Nolan: Mitigating object hallucinations in large vision-language models via dynamic suppression of language priors. *arXiv preprint arXiv:2602.22144*, 2026.

Sreetama Sarkar, Yue Che, Alex Gavin, Peter Anthony Beerel, and Souvik Kundu. Mitigating hallucinations in vision-language models through image-guided head suppression. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12481–12500, 2025.

Han Sun, Qin Li, Peixin Wang, and Min Zhang. Mitigating object hallucinations in lvlms via attention imbalance rectification. *arXiv preprint arXiv:2603.24058*, 2026a.

Haoxiang Sun, Tao Wang, Li Yuan, Jian Zhao, and Jiancheng Lv. From structure to synergy: A survey of vision-language perception paradigm evolution in multimodal large language models. *Information Fusion*, pp. 104285, 2026b.

Lexiang Tang, Xianwei Zhuang, Bang Yang, Zhiyuan Hu, Hongxiang Li, Lu Ma, Jinghan Ru, and Yuexian Zou. Not all tokens and heads are equally important: Dual-level attention intervention for hallucination mitigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9439–9447, 2026.

Changyao Tian, Danni Yang, Guanzhou Chen, Erfei Cui, Zhaokai Wang, Yuchen Duan, Penghao Yin, Sitao Chen, Ganlin Yang, Mingxin Liu, et al. Internvl-u: Democratizing unified multimodal models for understanding, reasoning, generation and editing. *arXiv preprint arXiv:2603.09877*, 2026.

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

Yujun Wang, Aniri, Jinhe Bi, Soren Pirk, and Yunpu Ma. Ascd: Attention-steerable contrastive decoding for reducing hallucination in mllm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10306–10314, 2026.

Sangmin Woo, Jaehyuk Jang, Donguk Kim, Yubin Choi, and Changick Kim. Ritual: Random image transformations as a universal anti-hallucination lever in large vision language models. *arXiv preprint arXiv:2405.17821*, 2024.

Jiulong Wu, Zhengliang Shi, Shuaiqiang Wang, Jizhou Huang, Dawei Yin, Lingyong Yan, Min Cao, and Min Zhang. Mitigating hallucinations in large vision-language models via entity-centric multimodal preference optimization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 19445–19461, 2025a.

Junfei Wu, Yue Ding, Guofan Liu, Tianze Xia, Ziyue Huang, Dianbo Sui, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. Sharp: Steering hallucination in lvlms via representation engineering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 14346–14361, 2025b.

Zongyu Wu, Yuwei Niu, Hongcheng Gao, Minhua Lin, Zhiwei Zhang, Zhifang Zhang, Qi Shi, Yilong Wang, Sike Fu, Junjie Xu, et al. Lanp: Rethinking the impact of language priors in large vision-language models. *arXiv preprint arXiv:2502.12359*, 2025c.

Hao Yin, Guangzong Si, and Zilei Wang. Clearsight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14625–14634, 2025.

Jianghao Yin, Qin Chen, Kedi Chen, Jie Zhou, Xingjiao Wu, and Liang He. Dynamic multimodal activation steering for hallucination mitigation in large vision-language models. *arXiv preprint arXiv:2602.21704*, 2026.

Haoran Zhou, Zihan Zhang, and Hao Chen. Extracting visual facts from intermediate layers for mitigating hallucinations in multimodal large language models. *arXiv preprint arXiv:2507.15652*, 2025.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations*, volume 2024, pp. 18378–18394, 2024.

A APPENDIX

A.1 SYNCONFACT DATASET CONSTRUCTION

We construct SynConFact to study how LVLMs respond when visual information conflicts with prior knowledge. The construction process consists of three stages: description generation, counterfactual image generation, and next-token prediction instance construction.

Description generation. We first prompt a large language model to generate 70 diverse counterfactual descriptions for each category. Each description specifies a familiar subject, object, or situation together with a counterfactual element that violates common prior knowledge. We consider six categories: object substitution, animal substitution, scene substitution, historical anachronism, material substitution, and attribute substitution. The generated descriptions are designed to cover diverse subjects and contexts while maintaining a clear conflict between the expected factual content and the intended counterfactual content.

Counterfactual image generation. The generated descriptions are provided to Image2 as image-generation prompts to produce the corresponding counterfactual images. We set the aspect ratio to (1:1) for all generated images to ensure a consistent image format across the dataset.

Next-token prediction instance construction. For each generated image, we use a multimodal language model to construct an incomplete sentence and paired factual and counterfactual candidates for next-token prediction. The factual candidates follow from the textual context and prior knowledge, whereas the counterfactual candidates describe the visual content. Both candidate sets are constrained to the same semantic slot. The target slots are objects, animal names, environments, functionally comparable historical items, materials or contents, and visual attributes for the six categories, respectively. To prevent textual leakage, the sentence must induce a natural factual continuation without revealing the counterfactual answer, and invalid instances are rewritten. Each instance is stored as a JSON object containing the image identifier, image path, incomplete sentence, and two candidate sets. The common prompt template, translated from the original Chinese instructions, and an example are provided below.

Prompt Used to Generate Dataset Instances

You are an assistant for constructing a counterfactual vision-language dataset. Given a counterfactual image from the specified category, generate high-quality JSONL data. Each line must contain exactly one JSON object with the following structure:

```
{"image_id":"000001",
 "image":"images/file.png",
 "text":"...",
 "factual_tokens":[...],
 "counterfactual_tokens":[...]}
```

Task type: [CATEGORY] mismatch.

Requirements:

1. The `text` field must be an incomplete sentence designed for next-token prediction.
2. The sentence must elicit an answer from the semantic slot defined for the selected category, rather than from an unrelated category.
3. The `factual_tokens` must naturally complete the sentence according to prior knowledge without considering the image.
4. The `counterfactual_tokens` must correctly describe the salient counterfactual content visible in the image.
5. The factual and counterfactual candidates must occupy the same semantic slot, such as object versus object, scene versus scene, or color versus color.
6. The sentence may identify the subject, action, function, or context needed to induce a natural factual continuation.
7. The sentence must not reveal the counterfactual content or contain clues that make it predictable without the image.
8. Do not use explicit prior cues such as "usually," "normally," "typically," "generally," "most likely," "common," or "expected."
9. The sentence must not be written as a question.
10. Rewrite the instance if the text alone reveals the counterfactual answer or fails to induce a natural factual continuation.
11. Output valid JSONL only, without explanations or Markdown code blocks.

Representative sentence patterns:

Object	`The [subject] is holding a`
Animal	`The animal [context] is a`
Scene	`The [subject/event] belongs in a`
Historical	`[Historical figure] is using a`
Material	`The [object/container] is filled with`
Attribute	`The [attribute] of the [subject] is`

Return JSONL only.

Example SynConFact Instance


```
{"image_id":"000011",
 "image":"images/example.png",
 "text":"The sauna room is occupied by a",
 "factual_tokens":
   ["person","man","human","bather"],
 "counterfactual_tokens":
   ["penguin","bird"]}
```




(a) Head 16 **(b)** Head 20 **(c)** Head 23 **(d)** Head 24

Figure 6: Head-wise attention maps in Qwen2-VL for the golf example.



(a) Head 2 **(b)** Head 4 **(c)** Head 5 **(d)** Head 14

(e) Head 16 **(f)** Head 17 **(g)** Head 18 **(h)** Head 23

Figure 7: Head-wise attention maps in Qwen2-VL for the chess example.

A.2 Additional Head-wise Attention Maps

To complement the analysis of representative heads, we provide additional attention maps for the golf and chess examples in Figs. 6 and 7, respectively. The visualizations show that attention heads focus on distinct regions of the image: some concentrate on the object or interaction relevant to the task, whereas others attend to the subject, background, or other visually salient regions.

A.3 Same-Index Head Correspondence

To validate the comparison between heads with the same index in our method, we measure the correspondence between individual head representations under the image-text and text-only conditions. For each sample, we consider only the first decoding step and extract the representations at the target prediction position from the final attention layer.

Table 6: Validation of same-index head correspondence across image-text and text-only conditions.

Model	Diag. Corr.	Self-Match Rate	Corr. Margin
Janus-Pro	0.810	99.69%	0.435
LLaVA-NeXT	0.644	95.39%	0.351
Qwen2-VL	0.618	94.79%	0.214
Qwen3-VL	0.775	99.43%	0.494

Suppose the evaluation set contains N samples and the model contains H attention heads. For sample n and condition $c \in \{\mathrm{IT}, \mathrm{T}\}$, let $\mathbf{x}_{n,h}^{c}$ denote the output of head h before the attention output projection, where IT and T denote the image-text and text-only conditions, respectively. We obtain the contribution of head h by applying its corresponding output projection block:

$$\mathbf{z}_{n,h}^{c} = \mathbf{W}_{h}^{O}\mathbf{x}_{n,h}^{c} \in \mathbb{R}^{d}, \qquad c \in \{\mathrm{IT}, \mathrm{T}\}, \tag{12}$$

where $\mathbf{W}_{h}^{O}$ is the output projection block corresponding to head h, and d is the model hidden size.

We first compute the mean of each head contribution across its hidden dimensions:

$$\bar{z}_{n,h}^{c} = \frac{1}{d}\sum_{r=1}^{d} z_{n,h,r}^{c}, \tag{13}$$

where r indexes the hidden dimensions.

The correspondence between image-text head h and text-only head k is measured using Pearson correlation:

$$C_{h,k}^{(n)} = \frac{\sum_{r=1}^{d}\left(z_{n,h,r}^{\mathrm{IT}} - \bar{z}_{n,h}^{\mathrm{IT}}\right)\left(z_{n,k,r}^{\mathrm{T}} - \bar{z}_{n,k}^{\mathrm{T}}\right)}{\sqrt{\sum_{r=1}^{d}\left(z_{n,h,r}^{\mathrm{IT}} - \bar{z}_{n,h}^{\mathrm{IT}}\right)^{2}}\sqrt{\sum_{r=1}^{d}\left(z_{n,k,r}^{\mathrm{T}} - \bar{z}_{n,k}^{\mathrm{T}}\right)^{2}}}. \tag{14}$$

Based on the resulting correlation matrix, we compute three correspondence metrics. The diagonal correlation measures the average correlation between heads with the same index:

$$\mathrm{DiagCorr} = \frac{1}{NH}\sum_{n=1}^{N}\sum_{h=1}^{H} C_{h,h}^{(n)}. \tag{15}$$

The self match rate measures how often the most correlated text-only head has the same index as the image-text head:

$$\mathrm{SelfMatch} = \frac{1}{NH}\sum_{n=1}^{N}\sum_{h=1}^{H}\mathbb{I}\left[\arg\max_{1\le k\le H} C_{h,k}^{(n)} = h\right] \times 100\%, \tag{16}$$

where $\mathbb{I}[\cdot]$ denotes the indicator function.

The correlation margin measures the advantage of the correlation with the head at the same index over the strongest correlation with any head at a different index:

$$\mathrm{CorrMargin} = \frac{1}{NH}\sum_{n=1}^{N}\sum_{h=1}^{H}\left(C_{h,h}^{(n)} - \max_{\substack{1\le k\le H\\ k\neq h}} C_{h,k}^{(n)}\right). \tag{17}$$

As shown in Table 6, the diagonal correlations range from 0.618 to 0.810 across the four LVLMs. Moreover, the self match rates range from 94.79% to 99.69%, and all models exhibit positive correlation margins. These results indicate that image-text heads are consistently aligned with their text-only counterparts at the same index, supporting the comparison of matching head indices when measuring semantic shifts induced by the image.

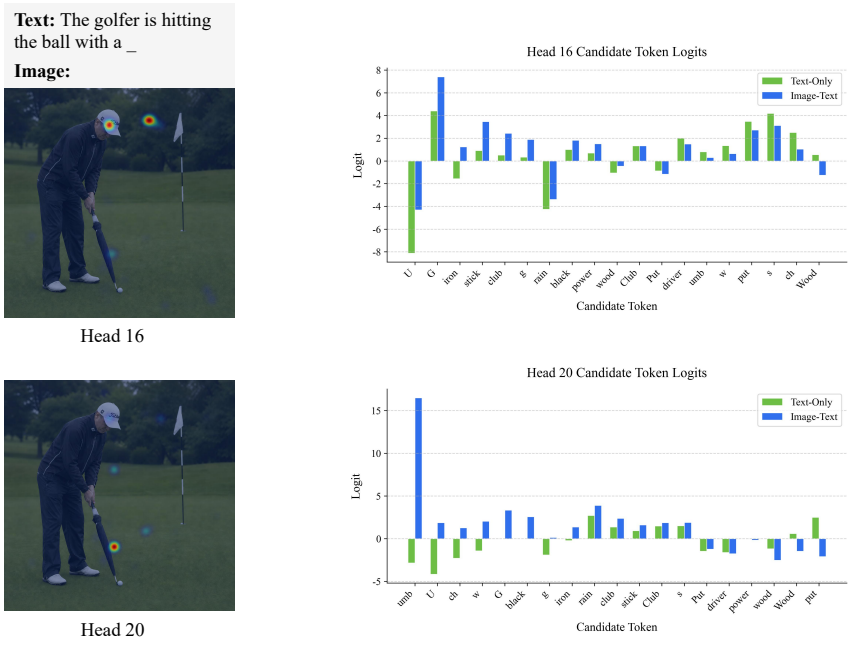


Figure 8: Image-induced semantic shifts across attention heads in Qwen2-VL for the golf example.

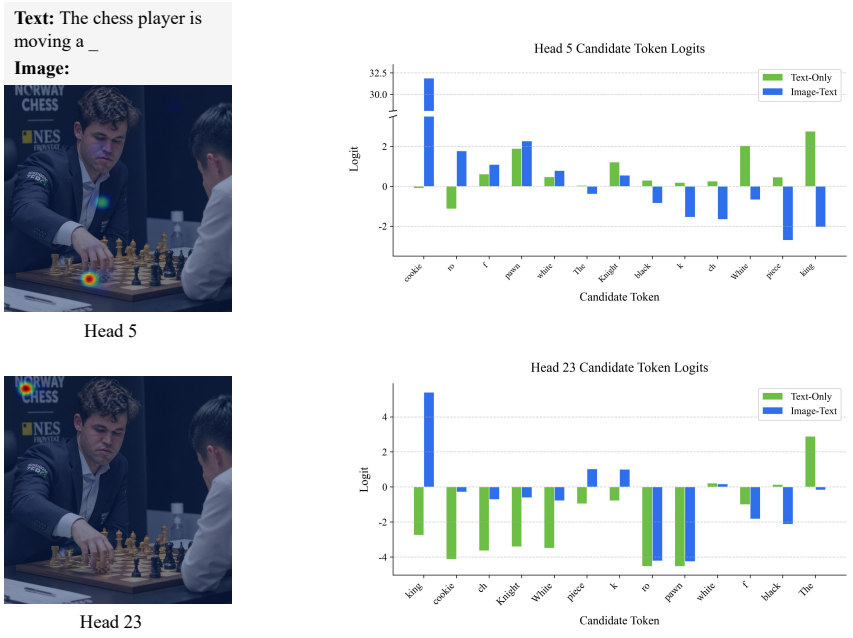


Figure 9: Image-induced semantic shifts across attention heads in Qwen2-VL for the chess example.

A.4 Extended Analysis of Representative Attention Heads

As an extension of the analysis presented in the main text, we provide more complete visualizations for two examples, as shown in Figs. 8 and 9. These visualizations show that the change from the text-only to the image-text predictive distribution over the candidate token set varies across heads, revealing distinct semantic shifts.

Table 7: Detailed evaluation results on the MME benchmark.

Category	Subtask	Janus-Pro		Qwen2-VL	
		Baseline	Ours	Baseline	Ours
Perception	Existence	175.0	175.0	195.0	195.0
	Count	136.7	136.7	155.0	155.0
	Position	**150.0**	148.3	**168.3**	163.3
	Color	145.0	**160.0**	185.0	185.0
	Posters	150.3	**152.0**	**184.7**	183.0
	Celebrity	153.8	153.8	**149.4**	149.1
	Scene	160.8	**163.0**	160.8	**161.5**
	Landmark	150.5	**152.0**	187.8	**189.3**
	Artwork	116.5	**116.8**	153.5	**155.5**
	OCR	162.5	162.5	**155.0**	147.5
	Total	1501.1	**1520.1**	**1694.4**	1684.2
Cognition	Comm.	**113.6**	112.1	144.3	144.3
	Num.	45.0	45.0	**132.5**	117.5
	Trans.	65.0	65.0	192.5	192.5
	Code	57.5	**60.0**	152.5	152.5
	Total	281.1	**282.1**	**621.8**	606.8
Overall		1782.2	**1802.3**	**2316.2**	2291.0

A.5 Experimental Settings

We use Janus-Pro-7B (Chen et al., 2025), LLaVA-v1.6-Mistral-7B (Liu et al., 2024a), Qwen2-VL-7B-Instruct (Wang et al., 2024), and Qwen3-VL-8B-Instruct (Bai et al., 2025). In our experiments, all methods use greedy decoding with a maximum of 8 generated tokens. For WITNESS, head enhancement is applied only in the last language model layer. At each decoding step, we reconstruct the candidate token set, recompute the head-wise JS divergence scores, and select heads for enhancement. Contrastive decoding and the final candidate token restriction are also applied at each step. The text-only and image-text branches share the same generated response prefix throughout decoding. For VHR (He et al., 2025), we reinforce heads only in the last language model layer with an amplification factor of 2. Heads with sensitivity scores above the layer-wise median are selected, corresponding to approximately half of the heads. For CK-PLUG (Bi et al., 2026), we set $\alpha = 0.4$. All experiments are conducted on NVIDIA A800 (80 GB) GPUs.

A.6 Detailed Results on the MME Benchmark

To further assess the effect of our method on general multimodal capabilities, we report detailed MME results for the Perception and Cognition categories of Janus-Pro and Qwen2-VL in Table 7. Although our method is primarily designed to resolve conflicts between visual information and prior knowledge, these results help characterize its impact on broader multimodal performance.

A.7 Effect of the Candidate Token Mask

We further visualize how the candidate token mask prevents tokens associated with irrelevant visual objects from interfering with semantic shift estimation. As shown in Fig. 10, the factual image contains a visually salient umbrella on the ground. However, the umbrella is not used by the golfer to hit the ball and is therefore irrelevant to the target prediction. Without the mask, the token *umbrella* could affect the shift estimate, causing it to reflect general visual salience rather than information relevant to the prediction. The candidate token mask excludes *umbrella* in the factual case but retains it in the counterfactual case, where the golfer uses the umbrella to hit the ball. This reduces interference from irrelevant visual objects and improves the reliability of the semantic shift scores.

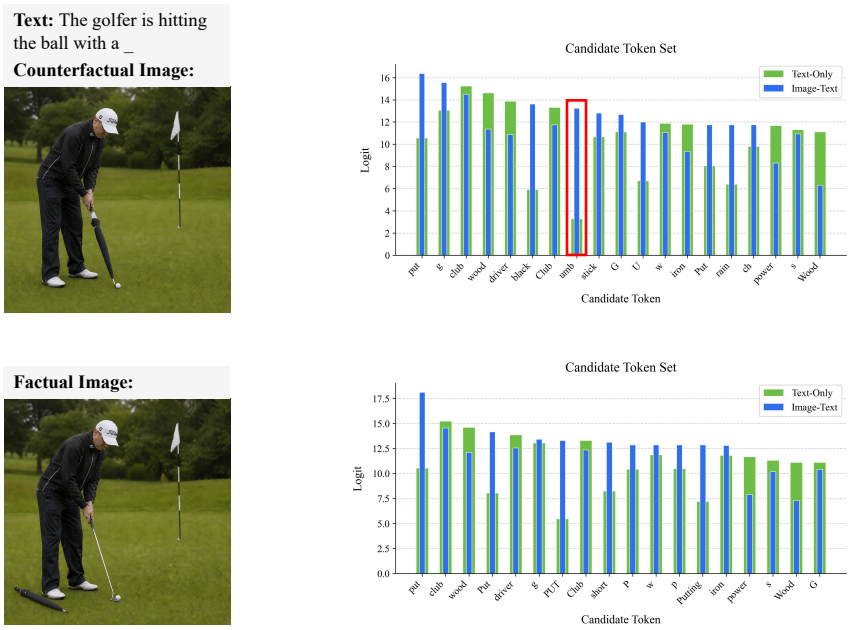


Figure 10: Candidate token set filtering prevents irrelevant counterfactual interference in Qwen2-VL.

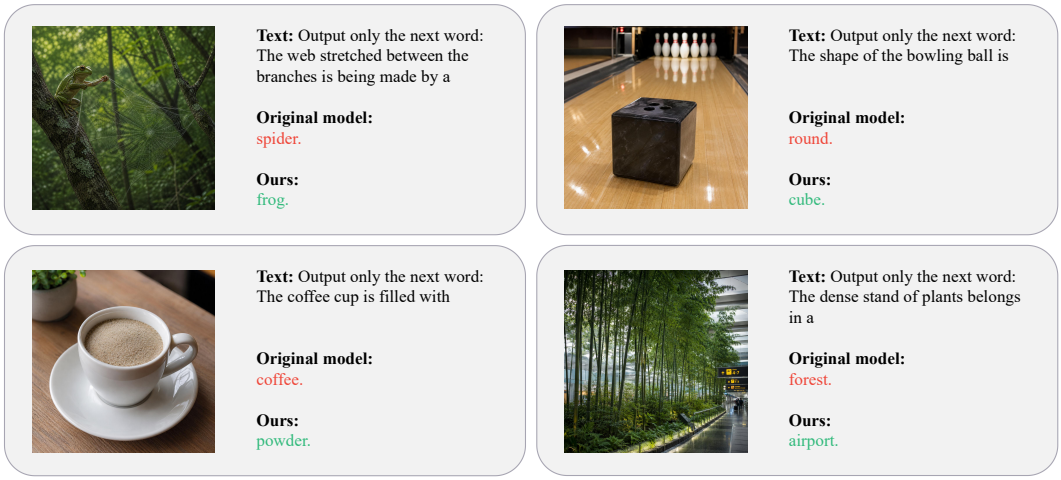


Figure 11: Case study on Qwen2-VL.

A.8 Case Study

Fig. 11 presents qualitative comparisons between the original Qwen2-VL model and Qwen2-VL with WITNESS on examples from different categories of SynConFact. The original model tends to generate next-token predictions dominated by familiar textual associations or prior knowledge, even when they conflict with the visual content. In contrast, Qwen2-VL with WITNESS produces predictions that are more consistent with the counterfactual information in the images. These examples illustrate that WITNESS strengthens the influence of visual information relevant to the task and reduces responses dominated by prior knowledge under different types of counterfactual conflicts.