Mitigating Action-Relation Hallucinations in LVLMs via Relation-aware Visual Enhancement

Anonymous ACL submission

Abstract

Large Vision-Language Models (LVLMs) have achieved remarkable performance on diverse vision-language tasks. However, LVLMs still suffer from hallucinations, generating text that contradicts the visual input. Existing research has primarily focused on mitigating object hallucinations, but often overlooks more complex relation hallucinations, particularly action relations involving interactions between objects. In this study, we empirically observe that the primary cause of action-relation hallucinations in LVLMs is the insufficient attention allocated to visual information. Thus, we propose a framework to locate action-relevant image regions and enhance the LVLM's attention to those regions. Specifically, we define the Action-Relation Sensitivity (ARS) score to identify attention heads that are most sensitive to action-relation changes, thereby localizing action-relevant image regions that contain key visual cues. Then, we propose the Relation-aware Visual Enhancement (RVE) method to enhance the LVLM's attention to these action-relevant image regions. Extensive experiments demonstrate that, compared to existing baselines, our method achieves superior performance in mitigating action-relation hallucinations with negligible additional inference cost. Furthermore, it effectively generalizes to spatial-relation hallucinations and object hallucinations.[1]

1 Introduction

Large Vision-Language Models (LVLMs) (Lu et al., 2024; Liu et al., 2024a; Chen et al., 2024d) have shown exceptional capabilities across various multimodal applications, ranging from visual question answering to complex reasoning. Despite these advancements, they are still prone to hallucinations, generating answers that are inconsistent with image content (Chen et al., 2024b; Kaul et al., 2024; Gunjal et al., 2024). This problem reduces the reliability of LVLMs in practical applications.

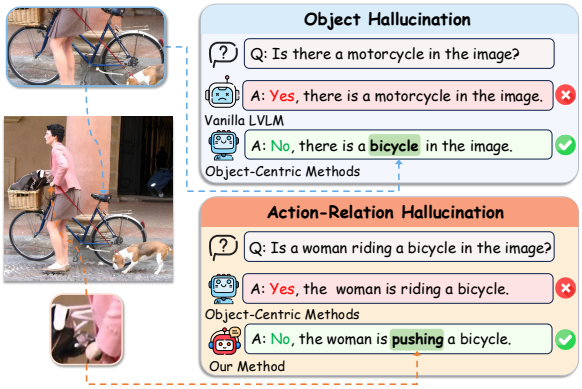

Figure 1: **Comparison between object hallucination and action-relation hallucination.** Existing object-centric methods effectively mitigate object hallucinations (top), but they fail to mitigate action-relation hallucinations (bottom).

Numerous methods have been proposed to mitigate hallucinations in LVLMs. Specifically, several studies adopt fine-tuning approaches, such as utilizing preference alignment training (Zhao et al., 2023; Sun et al., 2024) and training data refinement (Yu et al., 2024a). To avoid high training costs, other studies propose training-free methods, including modifying output logits (Chen et al., 2024c; An et al., 2025) and adjusting internal attention (Yin et al., 2025; Jiang et al., 2025). However, these training-free studies are primarily designed for object hallucinations and fail to address action-relation hallucinations. As illustrated in Figure 1, action-relation hallucinations involve not only objects but also complex interactions between them, requiring LVLMs to precisely capture action-relevant visual features. Please see Appendix A for further comparative analysis.

In this paper, we aim to mitigate relation hallucinations in LVLMs, with a particular focus on action relations. To this end, we first explore the

[1]The code and data will be released after acceptance.

reason behind action-relation hallucinations. We observe that despite image tokens constituting the vast majority of the input sequence, they receive disproportionately low attention compared to text tokens. Inspired by the above observation, we propose to enhance the LVLM's attention to action-relevant image tokens to mitigate action-relation hallucinations. Specifically, we propose a framework to first locate action-relevant image regions and then enhance the LVLM's attention towards these regions during inference.

To locate action-relevant image regions, we define the Action-Relation Sensitivity (ARS) score to measure the sensitivity of each attention head to action-relation changes. Visualization results verify that attention heads with high ARS scores explicitly focus on action-relevant image regions, thereby validating the effectiveness of the ARS score. Notably, we find that the middle layers have higher ARS scores than shallow and deep layers, indicating the middle layers are the most sensitive to action-relation changes.

Therefore, we propose the Relation-aware Visual Enhancement (RVE) method to enhance middle layers' attention toward action-relevant image regions, thereby mitigating action-relation hallucinations. Specifically, for each selected layer, we construct an enhancement mask based on attention heads with high ARS scores to locate action-relevant image regions. Besides, we observe that attention heads with low ARS scores often capture background noise common to all attention heads. Thus, we construct a denoising mask to locate these action-irrelevant image regions that should not be enhanced. Finally, we apply these masks to the attention maps of all heads within the selected layer to amplify focus on action-relevant regions.

Our main contributions are as follows:

- We define the ARS score to quantify the sensitivity of attention heads to action-relation changes. Leveraging this score, we reveal that middle layers are most sensitive to action-relation changes.
- We propose the training-free RVE method to enhance attention on action-relevant regions while reducing the interference of background noise.
- Extensive experiments demonstrate that our method can effectively mitigate action-relation hallucinations while generalizing to spatial-relation and object hallucinations, with negligible additional inference cost.

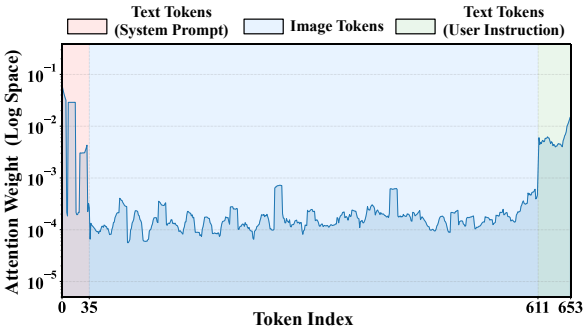

Figure 2: Attention weights (in the log space) of the last token allocated to the entire input sequence. Results reveal that the attention allocated to text tokens is approximately 10 to 100 times that of image tokens.

2 Preliminaries and Motivation

LVLM Generation. Given an input text $T = \{t_1, \ldots, t_{N_T}\}$ comprising N_T text tokens and an input image $I = \{v_1, \ldots, v_{N_I}\}$ comprising N_I image tokens, an LVLM employs a visual encoder followed by a modality connector to transform the input image tokens into visual embeddings $X_v \in \mathbb{R}^{N_I \times d}$, where d denotes the embedding dimension. For the text, a tokenizer converts the input text into text tokens, followed by an embedding module to obtain text embeddings $X_t \in \mathbb{R}^{N_T \times d}$. Subsequently, the LLM backbone processes the concatenated embeddings of X_v and X_t to autoregressively generate the next tokens.

Multi-Head Attention. The LLM backbone relies on the multi-head attention mechanism to integrate visual and textual information from input embeddings. For the h-th head (where $h = 1, \ldots, H$) in the l-th layer, we denote the full attention weights of the last token with respect to *all input tokens* as $W^{(l,h)} \in \mathbb{R}^{1 \times N}$, where $N = N_T + N_I$. We then obtain attention weights corresponding to the *image tokens*, denoted as $A^{(l,h)} \in \mathbb{R}^{1 \times N_I}$, and the corresponding pre-softmax attention scores $S^{(l,h)} \in \mathbb{R}^{1 \times N_I}$.

Motivation. To investigate the underlying causes of action-relation hallucinations, we analyze the attention distribution of LLaVA-1.5-7B over the entire input sequence during the generation process. As illustrated in Figure 2, our observations reveal a severe modality imbalance: the attention allocated to text tokens is approximately 10 to 100 times that allocated to image tokens. This suggests that the model may rely heavily on language priors while overlooking the visual information.

To further quantify such imbalance of attention weights, let $\mathcal{T}$ denote the set of text token indices

and $\mathcal{I}$ denote the set of image token indices. We quantify the ratio of attention allocated to text tokens and the ratio for vision tokens as follows:

$$r_{\text{att}}^{(l),t} = \frac{\sum_{i\in\mathcal{T}} W_{1,i}^{(l)}}{\sum_{j=1}^{N} W_{1,j}^{(l)}}, \quad r_{\text{att}}^{(l),v} = \frac{\sum_{i\in\mathcal{I}} W_{1,i}^{(l)}}{\sum_{j=1}^{N} W_{1,j}^{(l)}}, \tag{1}$$

where $W^{(l)} = \frac{1}{H}\sum_{h=1}^{H} W^{(l,h)}$. We also calculate the number ratio of text tokens and the number ratio of vision tokens as follows:

$$r_{\text{num}}^{t} = \frac{N_T}{N}, \quad r_{\text{num}}^{v} = \frac{N_I}{N}. \tag{2}$$

As illustrated in Figure 3, the attention allocated to image tokens is markedly lower than that given to text tokens, despite image tokens constituting the vast majority ($r_{\text{num}}^{v} = 87.8\%$) of the input sequence. Specifically, with the exception of the two layers closest to the input, the attention allocated to image tokens in the subsequent layers ranges from 4.2% to 16.4%, significantly lower than that allocated to text tokens. This further validates that during inference, the LVLM's focus on visual input is insufficient, leading to an output distribution biased towards language priors.

Based on these observations, we assume that the primary cause of action-relation hallucinations in LVLMs is the insufficient attention allocated to visual information. Therefore, we posit that **enhancing the LVLM's attention towards action-relevant image tokens is essential to mitigate action-relation hallucinations.** For instance, as illustrated in Figure 1, it is essential to enhance the attention specifically on the hand regions, which are highly relevant to the "pushing" action verb. To achieve this, we confront two key challenges:

Challenge 1: How to locate action-relevant image regions? Prior research has revealed that attention heads in LLMs exhibit functional specialization, such as retrieval heads for extracting relevant information from long contexts (Basile et al., 2025; Wu et al., 2024b). Inspired by these insights, we aim to identify the attention heads that are sensitive to action-relation changes. Subsequently, we locate the highly activated regions within these attention heads and select them as the action-relevant image regions. Please see Section 3.1 for details.

Challenge 2: How to effectively steer the LVLM's focus toward action-relevant image regions? We propose to amplify the attention weights corresponding to these action-relevant image regions, while reducing the interference of irrelevant background noise. This enables the model to focus more on action-relevant image information during inference, thereby mitigating action-relation hallucinations. Please see Section 3.2 for details.

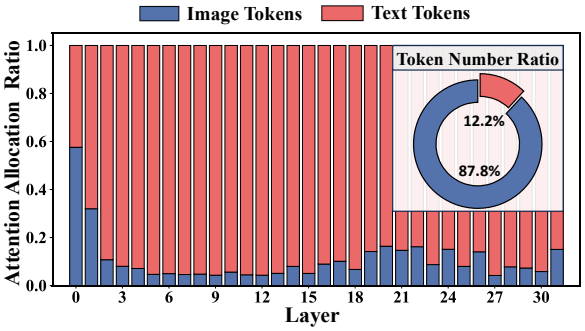

Figure 3: Analysis of the input token number ratios ($r_{\text{num}}^{t}, r_{\text{num}}^{v}$) and attention allocation ratios ($r_{\text{att}}^{t}, r_{\text{att}}^{v}$). While image tokens comprise 87.8% of the input sequence, their attention is disproportionately low.

3 Method

To address these two key challenges, we propose a training-free framework to mitigate action-relation hallucinations in LVLMs. As illustrated in Figure 4, the proposed framework comprises two modules: the action-relation-sensitive head identification module (Section 3.1) and the relation-aware visual enhancement module (Section 3.2).

3.1 Action-Relation-Sensitive Head Identification

To identify the attention heads that are sensitive to action relations, we first construct action-contrastive pairs. Specifically, given an input pair consisting of an image I and a text T, we generate a contrastive text $\hat{T}$ by substituting the action verb in T with a semantically distinct alternative. For example, given the input text "*Does a man hold a surfboard*," we change the verb "*hold*" to "*ride*," thereby creating a contrastive query, as shown in Figure 4. We utilize GPT-5 to automatically generate these contrastive samples.[1] For further details regarding the generation process and more examples, please see the Appendix B. Through this process, we construct input pairs (I, T) and $(I, \hat{T})$ that differ only in the verb. In this way, for the h-th head in the l-th layer, we obtain the visual attention weights $A^{(l,h)}$ and $\hat{A}^{(l,h)}$ corresponding to the inputs (I, T) and $(I, \hat{T})$, respectively.

To evaluate the sensitivity of attention heads to action-relation changes, we propose the following metric to quantify the divergence in attention distributions, which is termed the Action-Relation

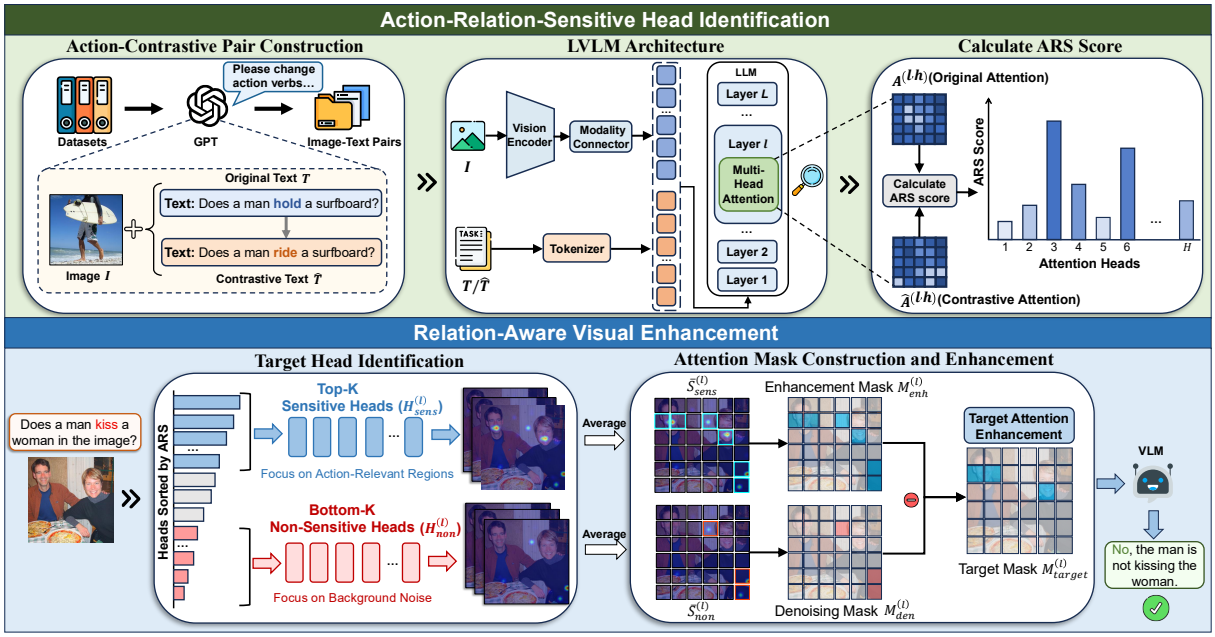

Figure 4: Overview of the proposed framework, consisting of two core components: (Top) Action-Relation-Sensitive Head Identification, which calculates ARS scores to identify attention heads critical for action-relation reasoning; and (Bottom) Relation-Aware Visual Enhancement, which leverages these attention heads to enhance action-relevant image regions during inference.

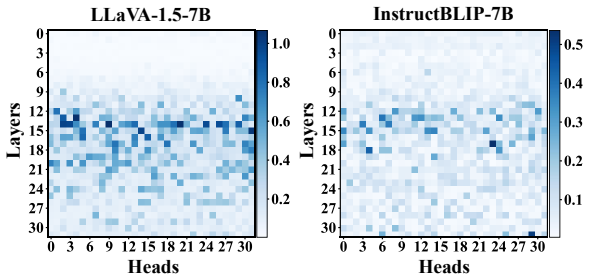

Figure 5: Distribution of ARS score values across layers and heads. Results indicate that the middle layers display high sensitivity to action-relation changes.

Sensitivity (ARS) score. Formally, the ARS score for the h-th head in the l-th layer is defined as:

$$\mathrm{ARS}^{(l,h)} = \frac{\|A^{(l,h)} - \hat{A}^{(l,h)}\|_F}{\frac{1}{2}\left(\|A^{(l,h)}\|_F + \|\hat{A}^{(l,h)}\|_F\right)}, \quad (3)$$

where $\|\cdot\|_F$ denotes the Frobenius norm, and the denominator is introduced for normalization. Figure 5 visualizes the distribution of ARS scores across all layers and attention heads. Results demonstrate a layer-wise pattern: attention heads within the middle layers exhibit notably higher ARS scores, whereas attention heads in shallow and deep layers display significantly lower scores. This observation suggests that the middle layers are more sensitive to action-relation changes.

Verifying the effectiveness of ARS scores. We visualize the attention maps of various heads within middle layers. If attention heads with high ARS scores indeed focus more on action-relevant image regions, then we can conclude that the ARS score effectively reflects the sensitivity to action-relation changes. Figure 6 displays the attention maps of the top-3 heads with the highest ARS scores and the bottom-3 heads with the lowest ARS scores in the middle layer. Results show that attention heads with the highest ARS scores focus on the image regions of the hand, which correspond to the action-relevant word "*point*." In contrast, the heads with the lowest ARS scores tend to focus on irrelevant regions. Such qualitative analysis verifies that the ARS score effectively reflects the sensitivity of attention heads to action-relation changes. Please see Section 4.2 for more verification of the effectiveness of ARS scores.

3.2 Relation-Aware Visual Enhancement

Building on the previous analysis, we propose the RVE method to amplify the LVLM's attention towards action-relevant image regions by modifying the pre-softmax attention scores $S^{(l,h)}$. As illustrated in Figure 6 (top), the action-relevant regions are primarily identified by the action-relation-sensitive heads. We also observe that action-relation-sensitive heads capture not only

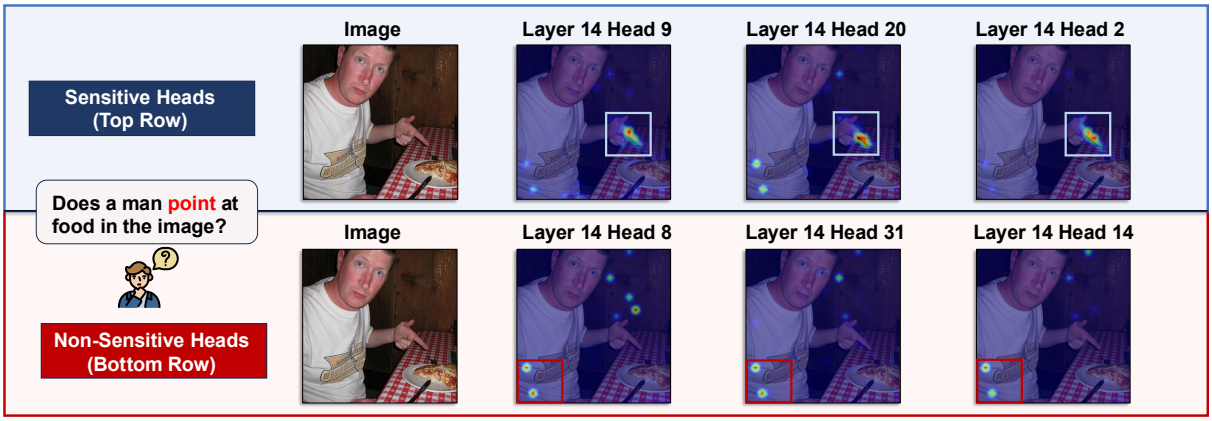

Figure 6: Visualization of attention maps at Layer 14 of LLaVA-1.5-7B. Sensitive heads (top) localize image regions relevant to the "point" action, while non-sensitive heads (bottom) focus on background noise, validating the effectiveness of ARS score.

action-relevant regions (e.g., *fingers*) but also action-irrelevant regions (e.g., *clothing*). Notably, these irrelevant regions are also captured by non-sensitive heads, as shown in Figure 6 (bottom). To avoid enhancing these irrelevant regions, we construct two attention masks to determine which image tokens require enhancement and which should be excluded.

Attention mask construction. Let $H^{(l)}_{\text{sens}}(K)$ denote the set of the top-K heads with the highest ARS scores, termed *sensitive heads*, and $H^{(l)}_{\text{non}}(K)$ denote the set of the bottom-K heads with the lowest ARS scores, termed *non-sensitive heads*. Based on these sets, we compute the averaged attention scores for the sensitive heads as $\bar{S}^{(l)}_{\text{sens}} = \mathbb{E}_{h \in H^{(l)}_{\text{sens}}}[S^{(l,h)}] \in \mathbb{R}^{1 \times N_I}$, and for the non-sensitive heads as $\bar{S}^{(l)}_{\text{non}} = \mathbb{E}_{h \in H^{(l)}_{\text{non}}}[S^{(l,h)}] \in \mathbb{R}^{1 \times N_I}$. Then, we construct an enhancement mask $M^{(l)}_{\text{enh}} \in \{0,1\}^{N_I}$ and a denoising mask $M^{(l)}_{\text{den}} \in \{0,1\}^{N_I}$ as follows:

$$M^{(l)}_{\text{enh}}[i] = \begin{cases} 1, & \text{if } \bar{S}^{(l)}_{\text{sens}}[i] \geq \tau_{\text{sens}} \\ 0, & \text{otherwise} \end{cases}, \quad M^{(l)}_{\text{den}}[i] = \begin{cases} 1, & \text{if } \bar{S}^{(l)}_{\text{non}}[i] \geq \tau_{\text{non}} \\ 0, & \text{otherwise} \end{cases}, \tag{4}$$

where τ_{sens} and τ_{non} are the values of the m-th largest elements in $\bar{S}^{(l)}_{\text{sens}}$ and $\bar{S}^{(l)}_{\text{non}}$, respectively. Here, $m = \lfloor \alpha \cdot N_I \rfloor$ is the number of selected tokens determined by the ratio α. In this way, the enhancement mask and the denoising mask are constructed to select the top-m image tokens in $\bar{S}^{(l)}_{\text{sens}}$ and $\bar{S}^{(l)}_{\text{non}}$, respectively.

To precisely locate action-relevant visual cues while avoiding the interference of background noise, we define the target mask $M^{(l)}_{\text{target}}$ as follows:

$$M^{(l)}_{\text{target}} = M^{(l)}_{\text{enh}} \odot (1 - M^{(l)}_{\text{den}}), \tag{5}$$

where $\odot$ denotes the element-wise multiplication.

As shown in Figure 4 (bottom), the enhancement mask $M^{(l)}_{\text{enh}}$ identifies both *the mouth regions* and *three background regions* as candidates for attention enhancement. Meanwhile, the denoising mask $M^{(l)}_{\text{den}}$ identifies these *three background regions* as noise that should not be enhanced. In this way, the target mask $M^{(l)}_{\text{target}}$ exclusively enhances the regions truly relevant to the action relation (i.e., *the mouth regions*) without amplifying action-irrelevant image regions.

Relation-aware visual enhancement. Finally, for each attention head h in layer l, we utilize the target mask $M^{(l)}_{\text{target}}$ to enhance the attention scores towards action-relevant image regions:

$$\tilde{S}^{(l,h)} = S^{(l,h)} + \beta \cdot (|S^{(l,h)}| \odot M^{(l)}_{\text{target}}), \tag{6}$$

where $\beta > 0$ is the enhancement coefficient.

4 Experiments

LVLMs and Benchmarks. We conduct experiments on **five LVLMs**: LLaVA-1.5-7B, LLaVA-1.5-13B (Liu et al., 2024b), LLaVA-NeXT-7B (Liu et al., 2024c), ShareGPT4V-7B (Chen et al., 2024a), and InstructBLIP-7B (Dai et al., 2023). To comprehensively evaluate the effectiveness and generalizability of our method, we conduct experiments under **four evaluation scenarios**: discriminative action-relation hallucination, generative action-relation hallucination, spatial-relation hallucination, and object hallucination. We compare our RVE method with **four**

Benchmark	Method	LLaVA-1.5-7B		LLaVA-NeXT-7B		InstructBLIP-7B		ShareGPT4V-7B		LLaVA-1.5-13B	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
MMRel† **(Real)**	Vanilla	71.12	77.32	70.58	77.21	65.47	74.32	72.96	78.44	74.51	79.33
	+ VCD	64.67	73.83	68.86	76.30	68.97	79.62	73.15	78.48	71.80	77.77
	+ ICD	73.30	78.54	70.79	77.34	61.10	71.99	72.31	78.07	74.06	79.01
	+ VAF	68.02	75.63	69.62	76.68	62.70	72.91	69.84	76.74	72.52	78.22
	+ **Ours**	**78.08**	**81.10**	**77.39**	**81.23**	**78.56**	**79.63**	**76.44**	**80.47**	**78.36**	**81.73**
MMRel (DALL-E)	Vanilla	69.06	73.28	67.54	74.67	69.65	75.19	69.06	74.86	72.44	77.55
	+ VCD	62.97	72.45	62.46	72.86	71.09	75.43	70.16	75.05	67.54	75.26
	+ ICD	69.57	72.35	67.62	74.79	67.29	73.44	68.81	74.53	71.77	77.19
	+ VAF	67.54	73.84	66.78	74.36	68.47	75.31	67.79	74.31	70.84	77.02
	+ **Ours**	**71.60**	**75.07**	**70.58**	**75.00**	**72.78**	**75.64**	**72.53**	**75.62**	**75.23**	**78.31**
R-Bench† **(Image)**	Vanilla	78.05	85.34	83.05	87.83	81.74	87.31	80.47	86.73	82.76	87.38
	+ VCD	78.62	85.54	82.50	87.49	81.35	86.87	81.40	86.98	81.36	86.23
	+ ICD	72.74	82.26	82.98	87.76	81.66	87.20	74.80	83.22	82.55	87.20
	+ VAF	77.69	85.23	82.89	87.76	81.38	87.24	79.81	86.38	82.89	87.53
	+ **Ours**	**79.10**	**85.64**	**83.50**	**88.13**	**82.48**	**87.59**	**81.84**	**87.45**	**83.26**	**87.76**
R-Bench (Instance)	Vanilla	62.94	73.45	64.59	83.15	66.70	77.57	66.37	76.54	69.12	79.37
	+ VCD	68.65	79.76	72.87	83.03	67.30	78.15	69.77	79.77	70.05	80.43
	+ ICD	59.22	69.83	68.50	78.80	68.53	79.33	66.25	76.40	69.01	79.30
	+ VAF	67.69	77.61	73.22	83.35	68.05	79.00	68.95	79.10	70.45	80.67
	+ **Ours**	**74.46**	**84.97**	**74.46**	**84.56**	**70.64**	**81.20**	**75.86**	**86.00**	**75.73**	**85.60**
AMBER (Relation)	Vanilla	58.41	45.43	62.62	53.86	68.87	75.26	67.13	61.61	73.44	71.92
	+ VCD	60.46	50.75	55.83	39.90	69.41	76.24	69.83	66.67	73.32	72.28
	+ ICD	59.38	47.43	61.06	50.91	68.39	77.09	68.09	63.15	73.20	71.63
	+ VAF	62.86	54.63	66.47	60.65	68.81	76.86	70.79	67.69	73.56	72.60
	+ **Ours**	**71.81**	**74.85**	**75.24**	**75.21**	**73.32**	**78.80**	**76.86**	**76.60**	**81.25**	**83.32**

† We focus our evaluation on the action-relation subsets for these benchmarks. For spatial-relation, please refer to Table 3.

Table 1: Performance comparison of vanilla LVLMs and different training-free mitigation methods on action-relation hallucinations across MMRel, R-Bench, and AMBER benchmarks. Please see Appendix F for case studies.

benchmarks: MMRel (Nie et al., 2024), R-Bench (Wu et al., 2024a), and AMBER (Wang et al., 2024a) for relation hallucination evaluation, and POPE (Li et al., 2023) for object hallucination evaluation. For each scenario, the specific benchmarks and their corresponding subsets are presented in Tables 1, 2, 3, and 4. For detailed descriptions of these subsets, please refer to Appendix C.

Baselines and Experimental Settings. We compare our method against vanilla LVLMs and three training-free mitigation methods (VCD (Leng et al., 2024), ICD (Wang et al., 2024b), and VAF (Yin et al., 2025)) applied to each vanilla model. Note that these three mitigation methods are originally designed for object hallucinations. To the best of our knowledge, there are currently no methods specifically designed for mitigating action-relation hallucinations. Regarding hyperparameters, we uniformly set $K = 5$ and $\beta = 1.0$ across all models. We set α to 0.05 for the LLaVA series and ShareGPT4V, and 0.5 for InstructBLIP. Please see Section 4.2 for detailed analysis.

Model	MMRel (Real)		MMRel (DALL-E)	
	Vanilla	Ours	Vanilla	Ours
LLaVA-1.5-7B	7.84	**7.91**	6.76	**6.89**
LLaVA-NeXT-7B	8.15	**8.21**	7.15	**7.32**
InstructBLIP-7B	6.19	**6.28**	7.41	**7.47**
ShareGPT4V-7B	8.18	**8.26**	6.87	**7.06**

Table 2: LLM-assisted open-ended evaluation on MMRel (Real) and MMRel (DALL-E).

4.1 Main Results

Performance on action-relation hallucinations. Results in Table 1 demonstrate that our method achieves consistent performance gains across all evaluated LVLMs and benchmarks, with *relative accuracy improvements* over the vanilla baseline ranging from 0.54% to 22.94%. Notably, our RVE method generalizes well across real and synthetic domains on MMRel, while maintaining effectiveness across different granularities on R-Bench. Furthermore, the consistent gains observed on both 7B and 13B models demonstrate the scalability and generalization of our RVE method.

Model	Method	MMRel		R-Bench	
		Real	DALL-E	Image	Instance
LLaVA-1.5-7B	Vanilla	50.25	50.41	78.21	60.85
	+ **Ours**	**54.61**	**54.05**	**79.08**	**75.19**
LLaVA-NeXT-7B	Vanilla	52.49	50.58	79.35	63.51
	+ **Ours**	**54.22**	**54.30**	**81.13**	**76.29**
InstructBLIP-7B	Vanilla	49.90	51.82	77.62	69.12
	+ **Ours**	**54.02**	**54.63**	**79.21**	**77.39**
ShareGPT4V-7B	Vanilla	53.01	53.31	79.67	64.61
	+ **Ours**	**57.35**	**55.37**	**80.31**	**77.30**

Table 3: Verifying the generalization of our RVE method to spatial-relation hallucinations.

Method	Random		Popular		Adversarial	
	Acc	F1	Acc	F1	Acc	F1
Vanilla	87.07	85.50	85.83	84.33	82.97	81.26
VCD	88.37	**87.91**	86.20	85.98	83.33	82.02
VAF	88.30	87.29	86.77	86.11	**83.72**	**83.05**
Ours	**88.63**	87.51	**87.17**	**86.59**	83.65	82.45

Table 4: Verifying the generalization of our RVE method to object hallucination. We report results on the three negative sampling settings of the POPE benchmark evaluated on MS-COCO, using LLaVA-1.5-7B.

Performance on open-ended generation tasks. In addition to discriminative metrics (accuracy and F1 scores), we evaluate the generative quality of the models using GPT-5-mini scoring (scale 0–10); please see Appendix D for details. As Table 2 shows, our method achieves higher quality scores across all architectures compared to the vanilla baseline. This demonstrates that our method effectively mitigates hallucinations without degrading the LVLM's generation capability.

Verifying generalization to spatial-relation hallucinations. As shown in Table 3, our method consistently outperforms the vanilla baseline across all LVLMs on spatial-relation hallucinations. Notably, our method achieves consistent performance gains with *relative improvements* over the vanilla baseline ranging from 0.80% to 23.57%, validating the generalization capability of our method to spatial-relation hallucinations.

Verifying generalization to object hallucinations. While our RVE method is designed to mitigate action-relation hallucinations, we further explore its generalizability to object hallucinations. As Table 4 shows, our method achieves competitive or superior performance on the POPE benchmark against the vanilla baseline and other object-hallucination mitigation methods. These findings effectively validate the generalization capability of our method in mitigating object hallucinations.

Strategy			LLaVA-1.5		InstructBLIP	
Global	Sens.	Denoise	Acc	F1	Acc	F1
			71.12	77.32	65.47	74.32
	✓		73.78	78.78	67.29	74.88
✓			74.88	79.37	69.47	74.65
	✓	✓	74.06	78.98	70.12	76.79
✓		✓	**78.08**	**81.10**	**78.56**	**79.63**

Table 5: Ablation of different enhancement scopes (global and sensitive-only) and the denoising mask on the real image subset of MMRel. See Appendix E for ablations on more models.

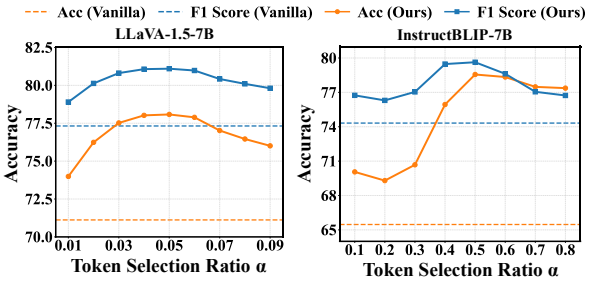

Figure 7: Impact of the selection ratio α on LLaVA-1.5-7B and InstructBLIP-7B evaluated on the real image subset of MMRel.

4.2 Ablation and Further Analysis

Impact of global enhancement. In Equation (6), we apply our RVE method to all attention heads in the target layer. In this experiment, we explore whether enhancing all heads (termed *Global*) is more effective than enhancing only the sensitive heads (termed *Sensitive-only*), where we select the top 50% of heads ranked by the ARS score. As Table 5 shows, under identical conditions, the *Global* enhancement consistently outperforms *Sensitive-only*. This indicates that aligning non-sensitive heads with sensitive ones further strengthens the LVLMs focus on action-relevant regions.

Impact of denoising mask. As Table 5 shows, our RVE method with the denoising mask consistently exhibits higher accuracy than the method without it. This proves the effectiveness of the denoising mask in preventing attention from being diverted to action-irrelevant regions.

Impact of hyperparameters. Figure 7 illustrates sensitivity to the selection ratio α across different architectures. For LLaVA-1.5-7B, the performance peaks at $\alpha = 0.05$, indicating that a small fraction of critical tokens suffices to capture action semantics given its large number of image tokens. In contrast, InstructBLIP-7B requires a significantly higher selection ratio, achieving optimal performance at $\alpha = 0.5$. We attribute this discrep-

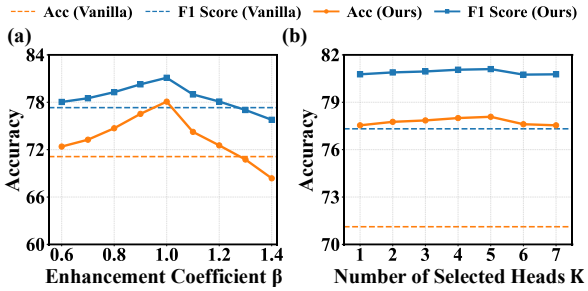

Figure 8: Results on LLaVA-1.5-7B. (a) Impact of the enhancement coefficient β on performance. (b) Impact of the number of selected heads K on performance.

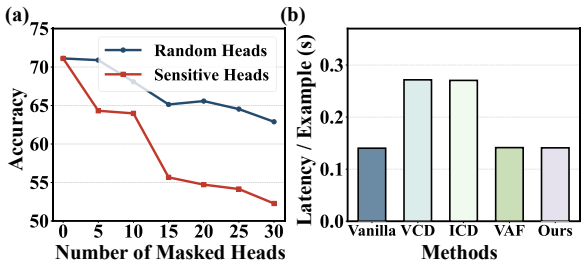

Figure 9: (a) Comparison of masking sensitive heads and random heads. (b) Inference latency per example across different methods.

ancy to the Q-Former compressing visual features into only 32 image tokens, thus requiring a larger retention proportion to preserve sufficient action information. Regarding the enhancement coefficient β, as shown in Figure 8 (a), performance steadily improves to a peak at $\beta = 1.0$. A gradual decline is observed thereafter, suggesting that excessive enhancement distorts the original feature distribution and diminishes the LVLM's attention to necessary contextual information. As Figure 8 (b) shows, performance reaches a peak when the number of selected heads $K = 5$. Notably, the performance remains stable across different values of K and consistently outperforms the vanilla baseline, validating the effectiveness of our ARS score in identifying action-relevant heads.

Verifying the effectiveness of selected sensitive heads. As Figure 9 (a) shows, masking randomly selected heads results in a slow performance decline, whereas masking the sensitive heads identified by the ARS score leads to a sharp accuracy drop. This significant gap confirms that the ARS score effectively locates the specific heads critical for action-relation understanding.

Inference speed. We further compare the inference latency of our method with baselines in Figure 9 (b). Methods like VCD and ICD double latency due to additional computational requirements. In contrast, our method maintains a speed comparable to the vanilla baseline, demonstrating that our method achieves effective hallucination mitigation with negligible extra inference cost.

5 Related Works

Large Vision-Language Models. LVLMs have achieved remarkable success in multimodal tasks (Caffagni et al., 2024; Li et al., 2025). These models typically comprise a vision encoder, a modality connector, and a pretrained LLM. Representative models employ diverse alignment strategies, including MLP-based projectors in LLaVA-1.5 (Liu et al., 2024b) and Q-Formers in InstructBLIP (Dai et al., 2023). Despite their impressive capabilities, these LVLMs still suffer from severe hallucinations, limiting their reliability in real-world applications.

Hallucination in LVLMs. Hallucination in LVLMs refers to inconsistencies between visual inputs and generated responses (He et al., 2025). Several approaches aim to mitigate hallucinations via reinforcement learning (Zhao et al., 2023; Yu et al., 2024b), data refinement (Yu et al., 2024a), or post-hoc revisers (Zhou et al., 2023). However, these methods incur high training costs. In contrast, training-free methods like OPERA (Huang et al., 2024) and Contrastive Decoding (e.g., VCD (Leng et al., 2024), ICD (Wang et al., 2024b)) offer lightweight solutions by adjusting attention or decoding distributions during inference. Nevertheless, existing methods mostly focus on mitigating object hallucinations, overlooking the more intricate issue of action-relation hallucinations (Zheng et al., 2025; Wu et al., 2024a). Consequently, we propose a training-free framework to mitigate action-relation hallucinations.

6 Conclusion

In this paper, we propose a training-free framework to mitigate action-relation hallucinations in LVLMs. Specifically, we define the ARS score to identify action-relation-sensitive attention heads and propose the RVE method to enhance attention toward action-relevant image regions. Extensive experiments across multiple benchmarks and LVLMs demonstrate that our method not only outperforms baselines in mitigating action-relation hallucinations but also effectively generalizes to spatial-relation and object hallucinations, with negligible additional inference cost.

Limitations

While our proposed framework effectively mitigates action-relation hallucinations with negligible additional inference cost, it exhibits certain limitations. First, since our method involves directly adjusting the LVLM's attention, it requires access to the LVLM's internal layers and representations. This restricts its applicability to closed-source models where only API-level access is available. Second, the optimal selection of layers for enhancement may vary across different model architectures and tasks to achieve optimal performance. Developing adaptive mechanisms that automatically determine the target layers based on dynamic metrics, such as generation confidence or attention entropy, remains a promising direction for future research.

Broader Impact and Ethics Statement

Our research focuses on mitigating action-relation hallucinations to enhance the reliability and truthfulness of LVLMs. We evaluate our method using publicly available datasets and LVLMs across multiple hallucination-related benchmarks. While our method demonstrates promising results, its effectiveness is constrained by the inherent capabilities of the base model, and improper usage may adversely affect the model's performance. To the best of our knowledge, our proposed method does not introduce additional ethical concerns regarding data privacy or social bias.

References

Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Ping Chen, Xiaoqin Zhang, and Shijian Lu. 2025. Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29915–29926.

Lorenzo Basile, Valentino Maiorca, Diego Doimo, Francesco Locatello, and Alberto Cazzaniga. 2025. Head pursuit: Probing attention specialization in multimodal transformers. *arXiv preprint arXiv:2510.21518*.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2024. The revolution of multimodal large language models: a survey. *arXiv preprint arXiv:2402.12451*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024a. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer.

Xiang Chen, Chenxi Wang, Yida Xue, Ningyu Zhang, Xiaoyan Yang, Qiang Li, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. 2024b. Unified hallucination detection for multimodal large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3235–3252.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. 2024c. Halc: object hallucination reduction via adaptive focal-contrast decoding. In *Proceedings of the 41st International Conference on Machine Learning*, pages 7824–7846.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024d. 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 (CVPR)*, pages 24185–24198.

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

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18135–18143.

Jinghan He, Kuan Zhu, Haiyun Guo, Junfeng Fang, Zhenglin Hua, Yuheng Jia, Ming Tang, Tat-Seng Chua, and Jinqiao Wang. 2025. 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)*, pages 3488–3501.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. 2025. 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.

Prannay Kaul, Zhizhong Li, Hao Yang, Yonatan Dukler, Ashwin Swaminathan, CJ Taylor, and Stefano Soatto. 2024. Throne: An object-based hallucination benchmark for the free-form generations of large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27228–27238.

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

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

Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. 2025. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges. *arXiv preprint arXiv:2501.02189*.

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

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

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

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. 2024. Deepseek-vl: Towards real-world vision-language understanding. *Preprint*, arXiv:2403.05525.

Jiahao Nie, Gongjie Zhang, Wenbin An, Yap-Peng Tan, Alex C Kot, and Shijian Lu. 2024. Mmrel: A relation understanding benchmark in the mllm era. *arXiv preprint arXiv:2406.09121*.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2024. Aligning large multimodal models with factually augmented RLHF. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13088–13110, Bangkok, Thailand. Association for Computational Linguistics.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. 2024a. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *Preprint*, arXiv:2311.07397.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024b. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. *arXiv preprint arXiv:2403.18715*.

Mingrui Wu, Jiayi Ji, Oucheng Huang, Jiale Li, Yuhang Wu, Xiaoshuai Sun, and Rongrong Ji. 2024a. Evaluating and analyzing relationship hallucinations in large vision-language models. In *International Conference on Machine Learning*, pages 53553–53570. PMLR.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. 2024b. Retrieval head mechanistically explains long-context factuality. *arXiv preprint arXiv:2404.15574*.

Hao Yin, Guangzong Si, and Zilei Wang. 2025. 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.

Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. 2024a. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12944–12953.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. 2024b. 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 (CVPR)*, pages 13807–13816.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. 2023. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*.

Kening Zheng, Junkai Chen, Yibo Yan, Xin Zou, Huiyu Zhou, and Xuming Hu. 2025. Reefknot: A comprehensive benchmark for relation hallucination evaluation, analysis and mitigation in multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6193–6212.

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

A Comparison of Difficulty between Object and Action-Relation Hallucinations

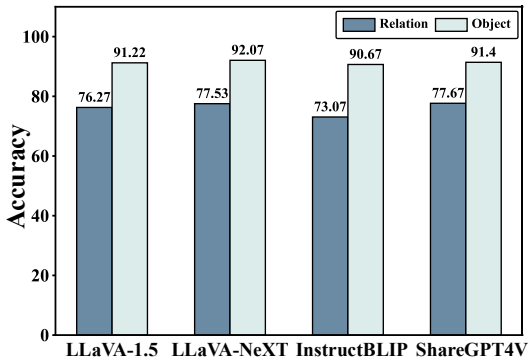

Figure 10: Performance comparison between object and action-relation hallucinations across four LVLMs.

To quantitatively assess the performance disparity between object and action-relation understanding, we conduct a comparative analysis based on the R-Bench (Wu et al., 2024a) dataset. We specifically utilize 1,500 images, each accompanied by paired queries designed to evaluate object hallucination and action-relation hallucination, respectively. This paired experimental design ensures that both tasks are evaluated under identical visual contexts, effectively eliminating potential biases arising from image complexity or content distribution. As illustrated in Figure 10, the results reveal a significant performance gap: while the accuracy on object queries consistently exceeds 90% across all models, the performance on action-relation hallucination lags behind by approximately 15%. This observation confirms that action-relation hallucination represents a more severe challenge that necessitates dedicated mitigation strategies.

B Generation of Action-Contrastive Image-Text Pairs

In this section, we describe the pipeline for generating action-contrastive image-text pairs.

B.1 Generation Pipeline

To construct valid contrastive pairs $(I, \hat{T})$ from the original pairs (I, T), we employ a systematic three-stage process leveraging the multimodal capabilities of GPT-5. The pipeline is structured as follows:

Stage 1: Initial Candidate Generation. We present the original image I and the instruction T to the model. As shown in Figure 11 (Prompt 1), the model is prompted to identify the body parts (e.g., hands) involved in the original action and propose a candidate verb v' associated with a different body part. For instance, a hand-related action such as "catch" may be replaced with a foot-related one like "kick".

Stage 2: Verification and Refinement. To verify the candidate verb generated in Stage 1, we submit it to a second validation pass (refer to Prompt 2 in Figure 11). This stage confirms whether the new verb strictly targets a distinct body part and remains syntactically compatible with the original context. If the candidate fails this verification, the model automatically regenerates a more suitable replacement.

Stage 3: Human Verification. To ensure the quality and reliability of the dataset, we manually review all generated pairs. This manual screening filters out instances that exhibit semantic ambiguity or remain linguistically unnatural after the automated stages. Only high-quality, unambiguous samples are retained for the final evaluation set.

B.2 Comprehensive Examples

Original Question	Original Verb	Contrastive Verb
Does a woman **hold** bread in the image?	hold	eat
Does a man **point** at food in the image?	point	look
Does a man **ride** a skateboard in the image?	ride	hold
Does a player **catch** a ball in the image?	catch	kick

Table 6: Action-contrastive samples generated by our pipeline.

Table 6 presents diverse examples of the generated contrastive samples from our dataset. As observed in the table, the original action verbs are replaced with plausible alternatives that involve different body parts (e.g., swapping the hand-related action "catch" with the foot-related action "kick"). This ensures semantic coherence in the generated text while maintaining a clear visual distinction between the image regions relevant to the contrastive verb and the original verb.

C Evaluation Benchmarks and Setup

In this section, we provide detailed descriptions of the four benchmarks used to evaluate our RVE method. For all evaluations, we employ greedy decoding and fix the random seed to 55 for all experiments to ensure the reproducibility of the results.

MMRel (Nie et al., 2024) is a large-scale relation-understanding benchmark that encompasses three

Prompt 1 (Initial Generation)

You are a visual action analysis assistant. Given an image and a question that contains an action word associated with a specific body part, identify the body part implied by the original action and propose a replacement action word that involves a different body part.

Output only the replacement action word (a single verb or a short verb phrase). Avoid actions that involve the same body part as the original.

Prompt 2 (Verification and Refinement)

Verify the proposed replacement action word.
(1) It must involve a body part different from the original action.
(2) It must be semantically appropriate in the context of the question.

If it satisfies both conditions, output it directly. Otherwise, output a new replacement action word that satisfies the rules. Output only the final action word.

Original question: {question_text}
Initial answer: {initial_answer}

Figure 11: The prompts used in our two-stage generation pipeline.

core relation categories (spatial, action, and comparative), supporting both discriminative and generative evaluations. It comprises approximately 22.5K question-answer pairs, consisting of around 15K discriminative Yes/No pairs and 7.5K open-ended questions. The benchmark features diverse image sources, including real-world images as well as synthetic domains generated by SDXL and DALL-E. In our experiments, we evaluate the RVE method on the real-image, DALL-E, and open-ended subsets.

R-Bench (Wu et al., 2024a) is a benchmark designed to evaluate relationship hallucinations in LVLMs. Constructed on the nocaps validation set, it utilizes a Yes/No verification format to specifically target inter-object relationships. R-Bench comprises two complementary subsets: image-level questions, which assess the existence of relationships within the global scene, and instance-level questions, which evaluate local visual understanding by explicitly grounding subjects and objects with colored bounding boxes or masks. The benchmark contains 11,651 curated questions in total, consisting of 7,883 image-level and 3,768 instance-level samples. For instance-level evaluation, models are prompted with specific templates (e.g., "*Is there {subject} in the red box...*") to verify precise relationship grounding.

AMBER (Wang et al., 2024a) is an LLM-free, multi-dimensional benchmark designed to systematically evaluate hallucinations across object existence, attributes, and relations. In this work, we specifically focus on the relation subset. Leveraging AMBER's high-quality human annotations, we adopt a discriminative QA format to induce relationship judgments using the specific prompt: "*Is there direct contact between the {object 1} and {object 2} in this image?*". We regard such contact-based relations as essential action primitives, as they represent the physical interaction between objects (e.g., *touching*).

POPE (Li et al., 2023) is a widely adopted benchmark for assessing object hallucinations in LVLMs. It evaluates models using Yes/No questions regarding object existence (e.g., "*Is there a {object} in the image?*"). The benchmark distinguishes three subsets based on the negative sampling strategy: Random, Popular, and Adversarial. Specifically, Random samples negative objects uniformly from the candidate set; Popular selects objects with the highest dataset frequency; and Adversarial targets objects that co-occur frequently with the ground-truth objects in the image. In this work, we conduct experiments specifically on the MSCOCO subset. This dataset comprises 3,000 test instances, consisting of 500 images with 6 questions per image.

D Details on the GPT-5-mini Evaluation

To comprehensively evaluate performance on open-ended tasks, we employ GPT-5-mini as an automated judge. The model scores each response by comparing it against the ground-truth answer. The specific prompt used for this automated evaluation is presented in Figure 12. The evaluator is instructed to focus specifically on the action relationship and the objects involved, assigning a score from 0 to 10.

Evaluation Prompt

We would like to request your feedback on the performance of an AI assistant in response to the user question displayed above. The user asks a question regarding the action relationship between two objects.

For your reference, the AI assistant is asked to answer with one sentence, which contains both objects and the action relationship.

Please rate the response from the AI assistant based on the ground-truth answer. Each response receives an overall score on a scale of 0 to 10, where a higher score indicates the AI response is more consistent with the ground-truth answer.

Please output the score for the Assistant.

Ground-truth answer: {ground_truth}
AI assistant response: {generated_answer}

Figure 12: The evaluation prompt used by the GPT-5-mini judge to evaluate open-ended generation tasks.

Strategy			LLaVA-NeXT-7B		ShareGPT4V-7B	
Global	Sens.	Denoise	Acc	F1	Acc	F1
			70.58	77.21	72.96	78.44
	✓		73.02	78.58	73.43	78.65
✓			74.38	79.38	75.88	80.14
	✓	✓	76.09	80.50	73.67	78.79
✓		✓	**77.39**	**81.23**	**76.44**	**80.47**

Table 7: Additional ablation study on LLaVA-NeXT-7B and ShareGPT4V-7B regarding enhancement scopes and the denoising mask.

E Additional Ablation Results

To further demonstrate the generalizability of our method, we provide the ablation results on LLaVA-NeXT-7B and ShareGPT4V-7B in Table 7. Consistent with the findings in the main text, the results confirm that: (1) Global enhancement outperforms the sensitive-only approach by effectively aligning non-sensitive heads; and (2) the denoising mask further boosts performance by avoiding the enhancement of irrelevant regions. Ultimately, the combination of these two components achieves the best results across different LVLMs.

F Case Studies

In this section, we present several case studies to evaluate model performance across multiple benchmark subsets. Specifically, we select examples from these three benchmarks, comprising seven diverse subsets: the real-image, DALL-E-generated, and open-ended subsets of MMRel, the image-level and instance-level (Mask and Box) tasks of R-Bench, and the relation subset of AMBER. In these examples, we can observe that our RVE method maintains effectiveness across diverse subsets compared to the vanilla baseline.

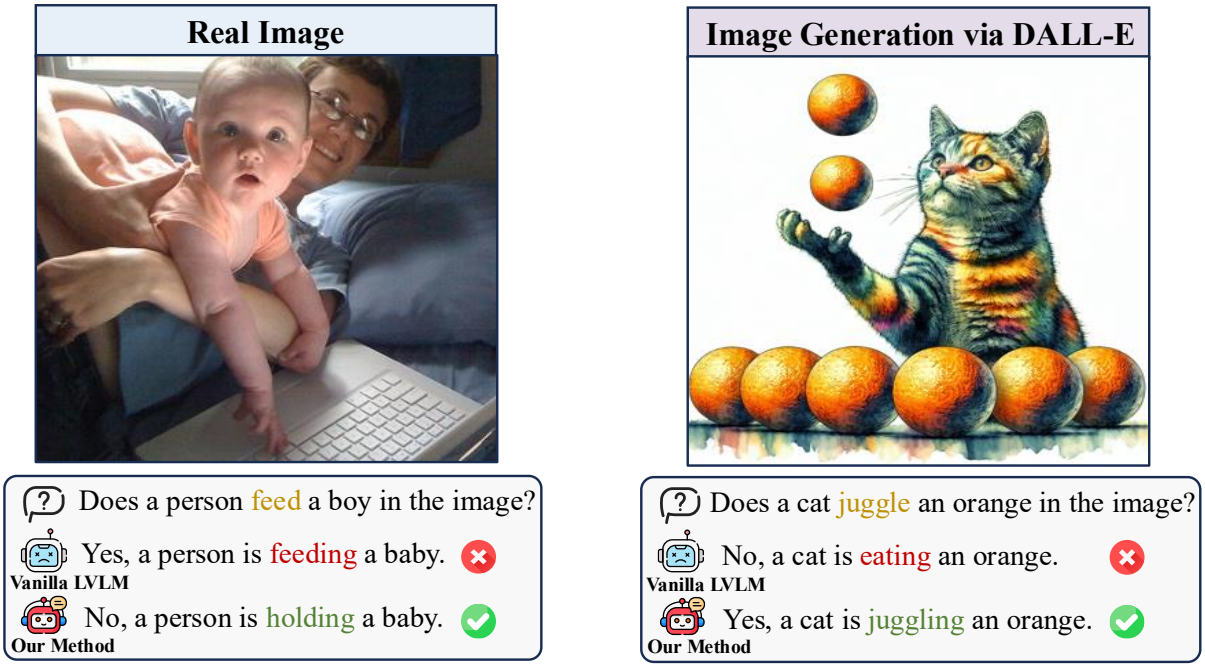

Figure 13: A case study of action-relation hallucinations on the MMRel benchmark across both real-world and DALL-E-generated scenarios.

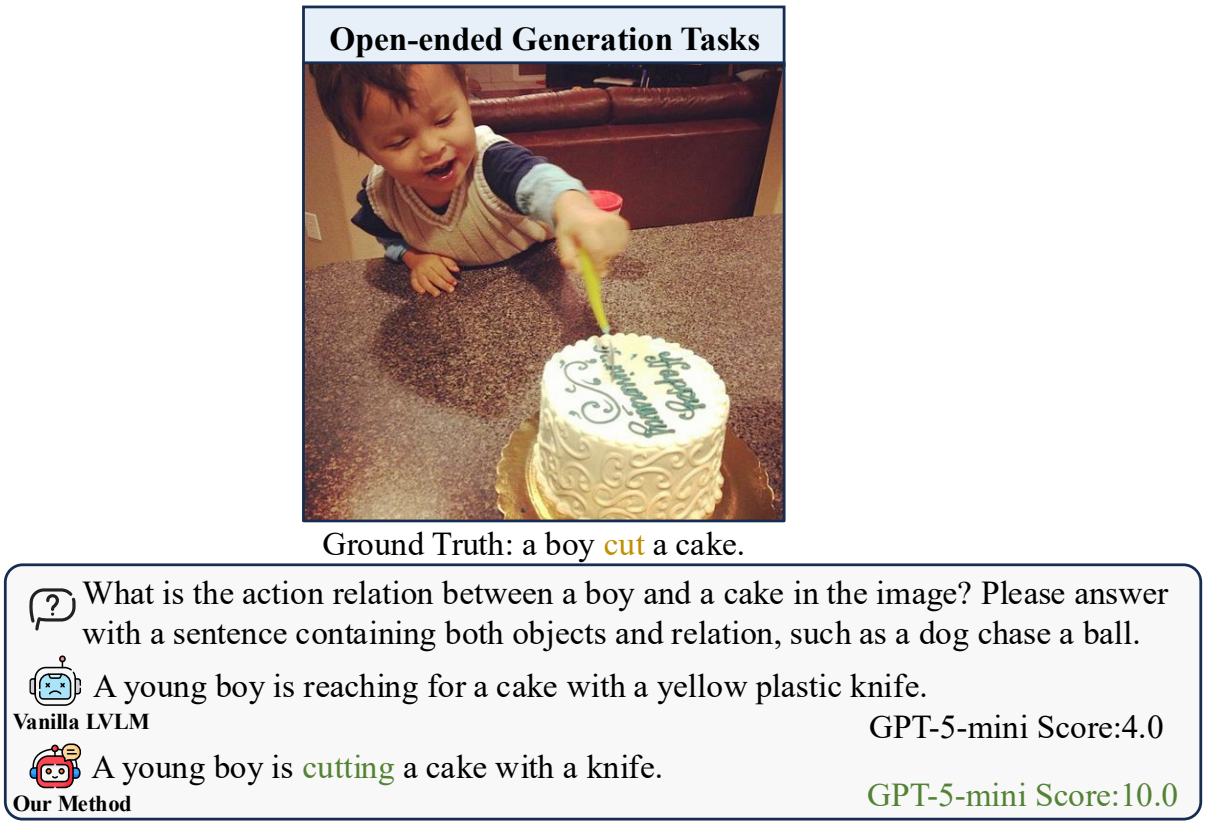

Figure 14: A case study of action-relation hallucinations in open-ended generation tasks on the MMRel benchmark.

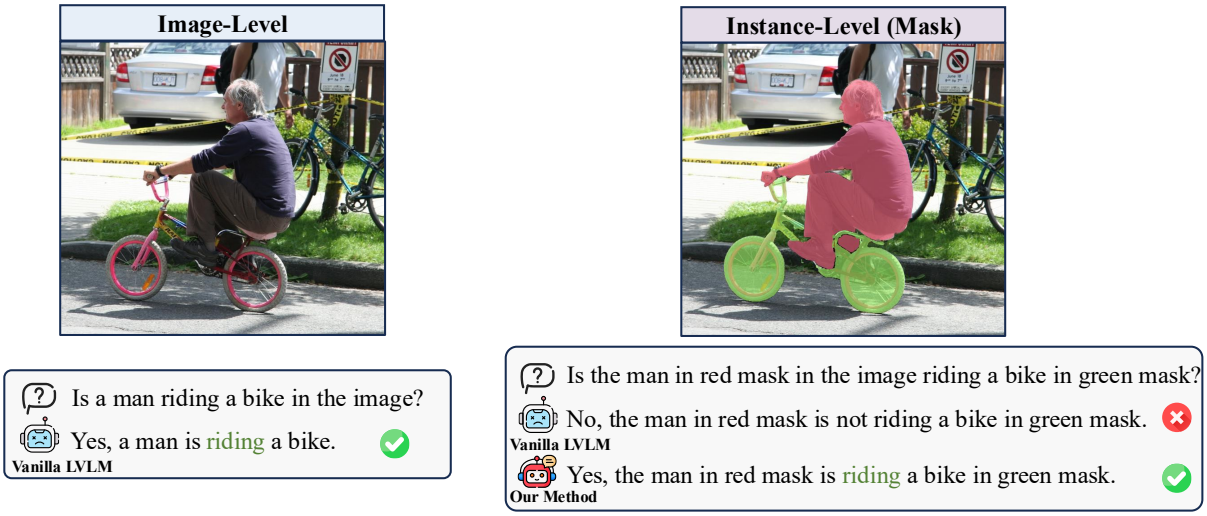

Figure 15: A case study comparing action-relation hallucinations between image-level and instance-level (Mask) tasks on the R-Bench benchmark.

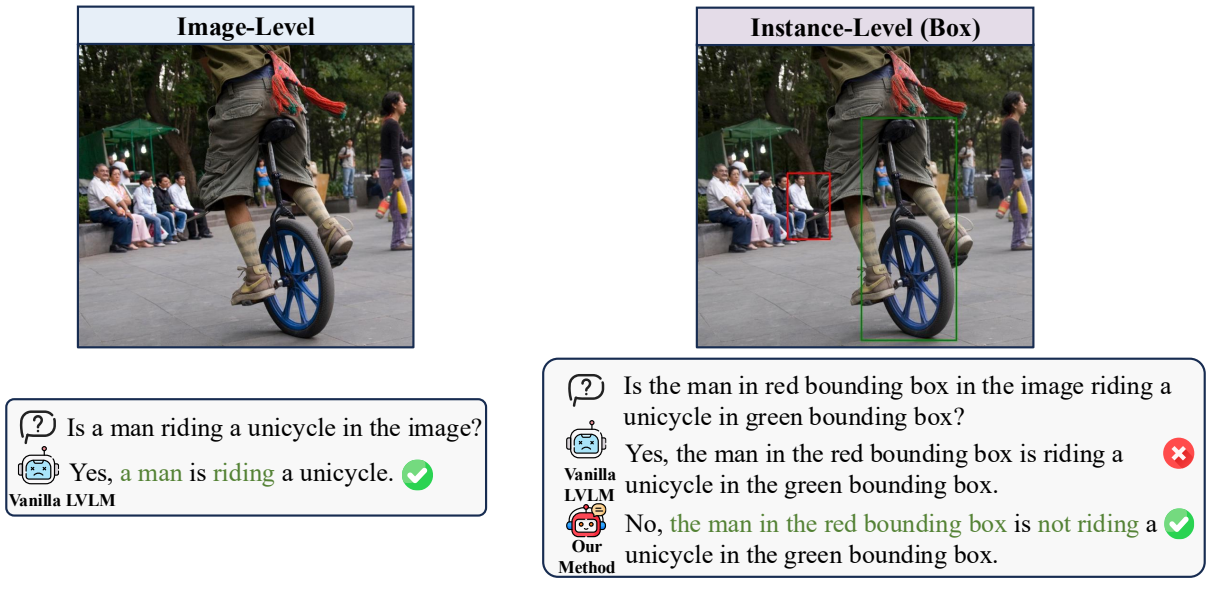

Figure 16: A case study comparing action-relation hallucinations between image-level and instance-level (Box) tasks on the R-Bench benchmark.

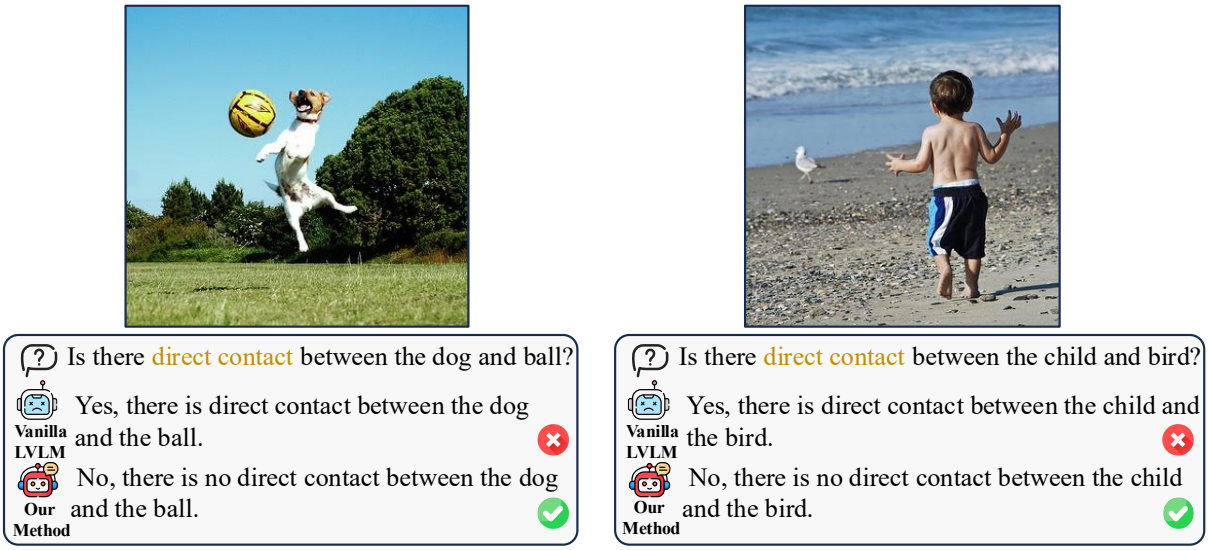

Figure 17: A case study of action-relation hallucinations on the AMBER benchmark.