Mitigating Hallucination in Large Vision-Language Models through Aligning Attention Distribution to Information Flow

Jianfei Zhao[1,2], Feng Zhang[1], Xin Sun[1,*], Chong Feng[1,3,*]
[1]School of Computer Science and Technology, Beijing Institute of Technology
[2]Zhongguancun Academy
[3]Southeast Academy of Information Technology, Beijing Institute of Technology
{zhqingan, bit_zhangfeng, sunxin, fengchong}@bit.edu.cn

Abstract

Due to the unidirectional masking mechanism, Decoder-Only models propagate information from left to right. LVLMs (Large Vision-Language Models) follow the same architecture, with visual information gradually integrated into semantic representations during forward propagation. Through systematic analysis, we observe that the majority of the visual information is absorbed into the semantic representations. However, the model's attention distribution does not exhibit sufficient emphasis on semantic representations. This misalignment between the attention distribution and the actual information flow undermines the model's visual understanding ability and contributes to hallucinations. To address this issue, we enhance the model's visual understanding by leveraging the core information embedded in semantic representations. Specifically, we identify attention heads that focus on core semantic representations based on their attention distributions. Then, through a two-stage optimization paradigm, we propagate the advantages of these attention heads across the entire model, aligning the attention distribution with the actual information flow. We evaluate our method on three image captioning benchmarks using five different LVLMs, demonstrating its effectiveness in significantly reducing hallucinations. Further experiments reveal a trade-off between reduced hallucinations and richer details. Notably, our method allows for manual adjustment of the model's conservativeness, enabling flexible control to meet diverse real-world requirements.[1]

1 Introduction

Large Vision-Language Models (LVLMs) (Dai et al., 2023; Bai et al., 2023; Liu et al., 2024c) integrate Large Language Models (LLMs) with visual encoders, aligning the extracted visual features with the semantic space of the LLMs to enable comprehensive understanding of visual content. However, they often suffer from hallucination (Wang et al., 2023b; Li et al., 2023; Liu et al., 2024b), where the models may generate content that is inconsistent with the visual evidence, thereby severely limiting their reliability in real-world scenarios.

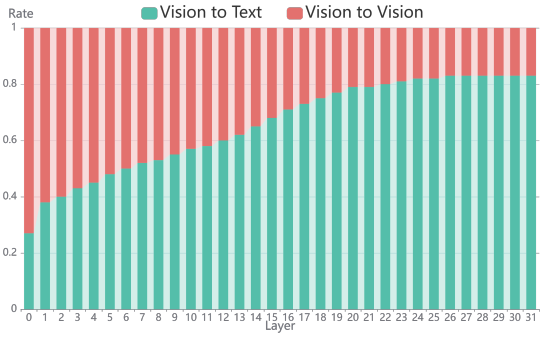

Figure 1: Visual information flow in the image captioning task on LLaVA-1.5. *Vision to Vision* and *Vision to Text* respectively denote the visual features' contributions to the visual representations and semantic representations.

Recent studies have explored methods to mitigate hallucinations in LVLMs. Leng et al. (2024) and Zhang et al. (2025) attribute hallucinations to the influence of language priors, proposing contrastive decoding-based methods to suppress language priors. However, they do not explicitly enhance the model's visual understanding capabilities. Other studies (Yin et al., 2025; Kang et al., 2025; He et al., 2025) argue that LVLMs exhibit biased mechanisms in visual attention distribution, and propose optimizations such as increasing the relative weight of key visual tokens. In this work, however, we point out that these approaches make suboptimal adjustments to the information flow, as they overlook the influence of information aggregation.

*Corresponding Authors.
[1]Code is available at https://github.com/beta-nlp/SEVI.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 24849–24863
November 4-9, 2025 ©2025 Association for Computational Linguistics

In fact, the unidirectional masked generation process in Transformer-based models can be regarded as a form of information flow, where information from earlier tokens in the input or generated sequence flows toward later tokens. Since LVLMs adopt this architecture, their understanding of visual content can be modeled as a process in which visual information flows into the semantic representations (since visual tokens are positioned before the textual tokens). Based on this perspective, we follow the Attention Rollout (Abnar and Zuidema, 2020) to conduct an in-depth analysis of information flow within LVLMs. The visualization results are shown in Fig. 1.

Specifically, we use $\mathbf{F}^l_{i,j}$ to represent the contributions of the i-th input embedding to the j-th representation in layer l. Considering the residual connection in the forward propagation, we formulate the recursive relation as $\mathbf{F}^l = (\mathbf{I}+\mathbf{W}_{\text{attn}})/2 \cdot \mathbf{F}^{l-1}$, where $\mathbf{W}_{\text{attn}}$ is the attention weight after average pooling across multi-heads. The contributions of each input embedding were normalized to 1 across all representations. We perform layer-wise quantification of visual embeddings' contributions to the visual and semantic representations, respectively.

As shown in the visualization, visual information progressively flows into semantic representations across layers. In the final layers, the majority of the visual information is integrated into the semantic representations. This indicates that **during the forward pass of the LVLMs, visual information is gradually encoded into the semantic representations**. While prior works largely concentrate on optimizing the attention distribution to visual features, they neglect the critical insight that essential information has already been integrated into semantic representations, where attention refinement may be more impactful.

To further validate this insight, we mask all visual representations and compare the consistency between the logits distribution of the vision-masked input and that of the regular (unmasked) input, as illustrated in Fig. 2. We observe that applying masking to visual representations starting from earlier layers leads to significant deviations in the model's outputs, whereas masking from later layers has minimal impact on the final outputs. This finding further supports our view that visual information has already been integrated into the semantic representations in the later layers.

Building upon the above findings, as the number of layers increases, textual representations progressively integrate more visual information and thus play an increasingly important role. Consequently, LVLMs should place greater attention on these integrated representations. To investigate this, we statistically analyze the attention allocation between semantic representations and others at each layer, as illustrated in Fig. 3. **The model's attention to semantic representations is significantly lower** than expected, accounting for only about one-fifth of the total attention. This phenomenon stems from the nature of supervision in LVLM training, which is primarily driven by soft supervision from end-to-end data and lacks explicit guidance on the attention distribution. We argue that **the model's attention distribution should align with the actual flow pattern of visual information**. Otherwise, hallucinations may arise, as the model does not sufficiently attend to regions enriched with visual information.

Figure 2: Jensen-Shannon divergence (JSD) between the regular and vision-masked logits distribution. *layer=l* represents the application of masking vision features starting from the l-th layer. Lower $\log(\text{JSD})$ denotes higher consistency between two distributions. The generation context is *The image depicts a _*.

To address the aforementioned challenge, we propose **SEVI** (Semantic-Enhanced Visual Interpretation), a novel training-free approach that augments the model's attention to those semantic representations that have absorbed visual information. Furtherly, we encourage the model to focus on the most meaningful representations in order to reduce the occurrence of hallucinations. Several studies (Wang et al., 2023c; Xiao et al., 2024) have revealed that contextual information tends to spontaneously concentrate in some tokens, forming anchor tokens that encapsulate core information. These tokens are typically characterized by receiving disproportionately high attention (Wang et al., 2023c). Therefore, our method focuses the model's

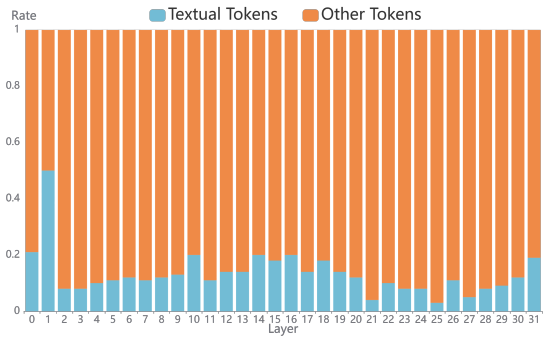

Figure 3: Attention distribution between semantic representations and all others, obtained by averaging across all attention heads.

attention on these core semantic representations at higher layers where information integration and aggregation occur. The core visual information contained within these representations provides more effective guidance for the model's understanding of visual inputs, thereby mitigating hallucination.

Specifically, we identify superior attention heads within the multi-head attention mechanism—those that attend to core semantic representations—and use their attention distribution as the target distribution to guide the optimization of the model's overall attention. To achieve this, we design a two-stage attention optimization paradigm, incorporating a smoothing mechanism to ensure stable performance improvements.

We evaluate the effectiveness of our method in mitigating hallucinations by testing it with five LVLMs: InstructBLIP (Dai et al., 2023), LLaVA-1.5 (Liu et al., 2024c), LLaVA-Next (Liu et al., 2024d), Qwen2-VL-Instruct (Wang et al., 2024b), and Qwen2.5-VL-Instruct (Bai et al., 2025), on three image captioning benchmarks: CHAIR (Rohrbach et al., 2018), AMBER (Wang et al., 2023a), and DetailCaps (Ye et al., 2025). Experimental results show that our approach significantly reduces hallucinations while supporting manual adjustment of the model's conservativeness, enabling a flexible trade-off between cautious outputs (fewer hallucinations) and comprehensive descriptions (richer details).

Our main contributions are as follows:

- We reveal that visual information is indeed integrated into semantic representations. However, the model's attention distribution does not align with this flow pattern, leading to hallucinations.
- We enhance the visual understanding capabilities of LVLMs by leveraging semantic representations. Specifically, we guide the model's attention toward core semantic representations through a two-stage optimization paradigm, effectively reducing hallucinations.
- We evaluate our method on the image captioning task, demonstrating its effectiveness in significantly reducing hallucinations across three benchmark datasets and five LVLMs. We observe a trade-off between reducing hallucinations and preserving detail, with our method enabling controllable conservativeness for task-specific needs.

2 Related Work

Various approaches have been proposed to mitigate hallucinations in LVLMs, such as improving training data quality (Liu et al., 2024a; Sun et al., 2024; Gunjal et al., 2024) or designing specific data formats (Hu et al., 2023; Zhai et al., 2024; Chen et al., 2025) to enhance model reliability. However, these training-based methods often suffer from limited scalability.

In recent years, the rapid evolution of LVLM backbones has brought increasing attention to training-free methods. Early approaches (Yin et al., 2024; Wang et al., 2024a; Zhou et al., 2024) primarily relied on post-processing techniques to correct hallucinated content after generation. However, the complexity of such pipelines and their dependence on external modules limit their practical usability. Researchers have since shifted focus toward investigating the root causes of hallucinations, exploring targeted optimizations in both the decoding strategies and the decoding processes. Some methods (Leng et al., 2024; Wang et al., 2024c; Zhang et al., 2025; Zhao et al., 2025) apply contrastive decoding algorithms to refine the logits, thereby reducing the likelihood of hallucination-related tokens. While these approaches improve the expressiveness of the model and enhance the transfer of visual information into textual form, they do not strengthen the model's intrinsic cross-modal understanding. Other works have proposed enhancements to the attention mechanism, such as correcting visual positional bias (Li et al., 2025; Zhu et al., 2025b), directly boosting visual attention (Zhu et al., 2025a; He et al., 2025), or redistributing attention weights (Kang et al., 2025; Yin et al., 2025). However, these methods primarily concentrate on visual attention and overlook the

vision-integrated semantic representation.

In contrast to these approaches, we analyze the information flow between the visual and textual modalities in LVLMs and focus the model's attention on core information within semantic representations, thereby reducing hallucinations more effectively.

3 Preliminaries

LVLMs consist of a visual encoder and an LLM backbone. High-dimensional features extracted by the visual encoder are concatenated with textual embeddings and jointly fed into the LLM. Subsequently, the visual and semantic features undergo cross-modal interaction through the Self-Attention layers within the LLM. The cross-modal interaction in layer l simplified as $\Delta \mathbf{X}^l = \text{softmax}(\mathbf{W})\mathbf{X}^l$, where $\mathbf{X}^l = [\mathbf{X}_V^l : \mathbf{X}_T^l]$ denotes the hidden states of visual features $\mathbf{X}_V^l$ and semantic features $\mathbf{X}_T^l$. $\mathbf{W} \in \mathbb{R}^{H \times L \times L}$ is the attention weights, H is the number of attention head and L is the model's context length.

When generating the (j+1)-th token, the model's attention to feature $\mathbf{X}_i^l$ at layer l across all heads is quantified by the weight value $\mathbf{W}_{i,j}$. A larger $\mathbf{W}_{i,j}$ indicates that the model's current layer relies more heavily on the information from $\mathbf{X}_i^l$ when determining the generated content. Therefore, we can intuitively infer that more important features $\mathbf{X}_i^l$ should generally correspond to larger attention weights $\mathbf{W}_{i,j}$.

4 Method

We attribute the occurrence of hallucinations to a misalignment between the attention distribution of LVLMs and the visual information flow. To address this, we enhance the visual interpretation capabilities of LVLMs by leveraging semantic representations that encapsulate visual information. We first identify superior attention heads that focus on core semantic representations, and then propagate their strengths to other heads through a two-stage optimization paradigm, as illustrated in Fig. 4.

4.1 Attention Distribution Alignment

Considering that visual information is progressively integrated into semantic representations through forward propagation—and that only a few of these representations aggregate core information—an intuitive approach is to increase the model's attention to these core semantic representations in the later layers. However, rigidly modifying the model's attention distribution may disrupt its inherent latent features, potentially resulting in suboptimal performance.

In the multi-head attention mechanism, attention heads naturally diversify after training, forming a heterogeneous and specialized structure (Vaswani et al., 2017). Accordingly, we identify superior attention heads that attend to core semantic representations and transfer their attention patterns to other heads, leading to a global refinement of the model's attention distribution.

Specifically, at each layer, we calculate the attention weight of each head with respect to the semantic representations. We then identify those heads that allocate more than 50% of their attention to semantic representations:

$$h \in \begin{cases} H_S & \text{if } \sum \mathbf{W}_S^h > \sum \mathbf{W}_O^h \\ H_O & \text{if } \sum \mathbf{W}_S^h \leq \sum \mathbf{W}_O^h \end{cases} \quad (1)$$

$\mathbf{W}_S^h$ and $\mathbf{W}_S^h$ respectively denote the attention weights of head h to semantic and other representations. H_S denotes the set of heads that focus more on semantic representations, whereas H_O denotes other heads. H_S are further categorized based on their attention distributions into *Core Semantic Heads* (H_{S_c}), which focus on core semantic representations, and *Global Semantic Heads* (H_{S_g}), which fail to effectively attend to them:

$$h \in \begin{cases} H_{S_c} & \text{if } \max \mathbf{W}_S^h > \kappa \cdot \sum \mathbf{W}_S^h \\ H_{S_g} & \text{if } \max \mathbf{W}_S^h \leq \kappa \cdot \sum \mathbf{W}_S^h \end{cases} \quad (2)$$

where κ is a hyperparameter and $H_{S_c} \cup H_{S_g} = H_S$ holds. We then take the core semantic heads to construct a target distribution to optimize the model's attention distribution.

To ensure stable information integration within the model, we propose a two-stage feature optimization paradigm that more naturally guides the model's understanding of visual information.

In the first stage, we use the global semantic heads to guide the other heads, aligning the model's attention transition to semantic representations and harmonizing with the visual information flow. In the second stage, we leverage the core semantic heads to guide the global semantic heads, encouraging the model to concentrate more on critical information and thereby suppressing hallucinations.

The model's internal attention is then optimized through a two-stage refinement process.

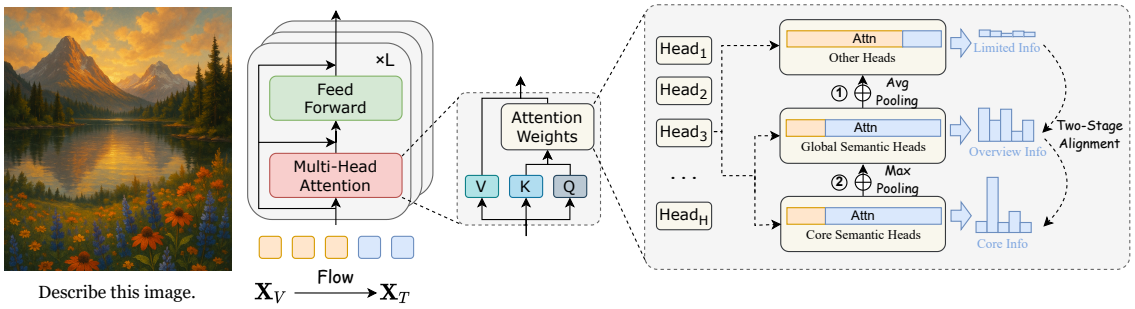

Figure 4: The diagram of the two-stage attention distribution alignment. We first categorize attention heads based on their focus on semantic representations into semantic heads and other heads. Semantic heads are further divided into core semantic heads and global semantic heads, depending on whether they attend to core semantic representations. We then align the model's attention distribution with the core semantic heads through a two-stage optimization process.

During this process, we incorporate a weighted smoothing mechanism to adjust the attention distribution of each head:

$$\hat{\mathbf{W}}^h = \frac{\mathbf{W}^h + \omega \mathbf{W}'}{1+\omega} \tag{3}$$

where $\hat{\mathbf{W}}^h$ denotes the optimized attention weights of head h, $\mathbf{W}'$ represents the target weight distribution derived from superior heads, and ω is a hyperparameter that controls the extent of the modification.

We first use the global semantic heads to guide the other heads, aiming to align the model's attention distribution with the flow of visual information. To this end, we apply average pooling across the global semantic heads to construct the target attention distribution:

$$\hat{\mathbf{W}}_O^h = \frac{\mathbf{W}_O^h + \omega \cdot \frac{1}{|H_{S_g}|} \sum_{t \in H_{S_g}} \mathbf{W}_{S_g}^t}{1+\omega}, \tag{4}$$

Subsequently, we use the core semantic heads to guide the global semantic heads, with the goal of enhancing the model's focus on critical information. Here, we employ max pooling across core semantic heads to form the corresponding target distribution:

$$\hat{\mathbf{W}}_{S_g}^h = \frac{\mathbf{W}_{S_g}^h + \omega \cdot \max_{t \in H_{S_c}} \mathbf{W}_{S_c}^t}{1+\omega} \tag{5}$$

We provide the algorithm of the two-stage optimization process in Appendix A.

4.2 Mitigating Aggravated Language Priors

We optimize the model's attention distribution to align with the flow pattern of visual information, which ultimately results in increased attention to semantic representations. However, while semantic representations incorporate visual information, they also inherently contain their own contextual features. As a result, increasing the model's reliance on semantic representations will exacerbate the issue of language priors (Wu et al., 2022; Ren et al., 2023; Chen et al., 2025).

To address this issue, we employ CICD (Zhao et al., 2025) to eliminate language priors while preserving visual information. CICD uses the cross-image consistency of language priors to identify detrimental priors and mitigate them by contrastive decoding:

$$\begin{aligned}\text{logit}(y_t \mid v, x, y_{<t}) =& (1+\alpha)\hat{\text{logit}}_\theta(y_t \mid v, x, y_{<t}) \\ & - \alpha \text{logit}_\theta(y_t \mid v', x, y_{<t})\end{aligned} \tag{6}$$

where $\hat{\text{logit}}_\theta$ is the logit distribution after attention distribution alignment and logit_θ is the regular logit distribution with a distinct image v'. α is defined as follows:

$$\alpha = 1 - \log_{10}(\text{JSD}(\hat{\text{logit}}_\theta; \text{logit}_\theta)) \tag{7}$$

where $\text{JSD}(\cdot;\cdot)$ represents Jensen-Shannon Divergence.

5 Experiments

5.1 Experimental Setup

Due to space constraints, we present only the key aspects of our experimental setup here. Detailed settings can be found in Appendix B.

Implementation Details Our method guides the model to focus on core information, thereby reducing the occurrence of hallucinations. In practice,

Method	LLaVA-1.5			InstructBLIP			LLaVA-1.5-13B		
	Cs↓	**Ci↓**	**R ↑**	**Cs↓**	**Ci↓**	**R↑**	**Cs↓**	**Ci↓**	**R ↑**
Maximum Generation Length is 64									
Regular	24.4	8.9	56.6	35.6	13.2	56.4	24.0	7.8	56.5
VCD (Leng et al., 2024)	25.0	8.3	59.0	32.2	10.3	60.6	23.4	7.5	59.4
ICD (Wang et al., 2024c)	23.2	8.1	58.4	29.8	9.8	60.6	18.8	6.6	58.4
PAI (Liu et al., 2024e)	20.0	6.2	56.9	26.0	8.9	53.9	17.4	5.3	58.7
IBD (Zhu et al., 2025a)	21.2	6.9	58.8	27.8	9.2	60.3	22.4	7.1	**59.7**
VAR (Kang et al., 2025)	25.2	8.6	55.3	-	-	-	25.8	8.5	56.3
AD-HH (Yang et al., 2025)	19.4	6.3	52.5	-	-	-	20.0	6.5	53.9
VAF (Yin et al., 2025)	26.2	9.3	56.6	32.0	12.0	55.3	23.2	8.0	57.6
DeGF (Zhang et al., 2025)	22.4	7.2	58.2	32.4	11.0	59.4	22.4	6.9	59.6
CICD (Zhao et al., 2025)	18.0	6.1	**59.6**	23.8	7.7	**62.2**	20.4	6.6	59.4
$\text{SEVI}_{\text{Balanced}}$	18.8	5.5	59.5	27.8	8.8	60.9	17.8	5.5	57.6
$\text{SEVI}_{\text{Focused}}$	**14.8**	**4.7**	54.1	**15.4**	**5.9**	50.9	**13.4**	**4.5**	54.4
Maximum Generation Length is 512									
Regular	54.6	16.4	72.6	62.6	19.5	66.9	58.8	17.0	73.4
VCD (Leng et al., 2024)	59.8	17.8	**75.6**	64.8	18.8	**71.9**	60.2	16.4	**76.9**
ICD (Wang et al., 2024c)	57.0	15.0	74.6	59.0	17.1	69.2	55.0	14.5	76.4
PAI (Liu et al., 2024e)	41.2	10.4	68.6	67.6	19.4	68.0	35.4	9.5	73.1
IBD (Zhu et al., 2025a)	57.6	16.5	74.2	57.6	15.7	70.8	50.6	14.3	76.1
VAR (Kang et al., 2025)	60.0	18.3	72.6	-	-	-	54.8	15.3	73.8
AD-HH (Yang et al., 2025)	46.6	12.6	66.5	-	-	-	50.0	13.5	71.0
VAF (Yin et al., 2025)	58.8	19.0	71.3	58.6	17.8	66.9	57.2	16.2	73.7
DeGF (Zhang et al., 2025)	57.4	16.3	75.5	59.0	17.7	71.5	51.8	14.2	75.5
CICD (Zhao et al., 2025)	43.8	11.7	75.0	49.8	13.7	70.3	44.4	12.9	76.0
$\text{SEVI}_{\text{Balanced}}$	34.8	9.0	68.3	41.0	12.4	67.1	28.6	8.4	68.7
$\text{SEVI}_{\text{Focused}}$	**17.8**	**5.5**	56.9	**18.8**	**8.4**	51.6	**15.0**	**5.1**	55.6

Table 1: Results on CHAIR. $\text{SEVI}_{\text{Balanced}}$ and $\text{SEVI}_{\text{Focused}}$ represent our method operating in the *Balanced* and *Focused* mode, respectively. **Cs** and **Ci** represent CHAIRs and CHAIRi, **R** represents Recall.

we observe a trade-off between hallucination suppression and the richness of generated details, influenced by the model's degree of attentional focus. To address this, we introduce hyperparameters to control the model's focus on the core information. Based on empirical observations, we design two sets of hyperparameter configurations: (1) *Focused* mode that aggressively minimizes hallucinations, and (2) *Balanced* mode that strikes a compromise between detail retention and hallucination reduction. Specifically, *Focused* mode optimizes attention distribution starting from the 5th layer with $\omega = 4$, while *Balanced* mode begins from the 9th layer with $\omega = 0.5$.

In addition, we set the parameter κ in Eq. 2 to 0.2. Further details are discussed in Appendix B.3. We employ sampling decoding for the next-token prediction with default settings. All experiments are performed on a single NVIDIA A800 40G GPU.

Benchmarks We evaluate the performance of our method across three widely adopted multimodal hallucination benchmarks on the image captioning task. These include CHAIR (Rohrbach et al., 2018), DetailCaps (Ye et al., 2025), and AMBER (Wang et al., 2023a).

Evaluated LVLMs To examine the generalizability of our approach, we apply it to five LVLMs drawn from three representative model families: InstructBLIP (Dai et al., 2023); two models from the LLaVA family (LLaVA-1.5 (Liu et al., 2024c) and LLaVA-Next (Liu et al., 2024d)); and two from the Qwen series (Qwen2-VL-Instruct (Wang

Method	LLaVA-1.5			InstructBLIP		
	CAPTURE↑	**Cs↓**	**Ci↓**	**CAPTURE↑**	**Cs↓**	**Ci↓**
Regular	52.60	55.7	17.4	52.99	58.6	18.0
VCD (Leng et al., 2024)	52.91	55.7	16.8	53.20	59.6	18.9
ICD (Wang et al., 2024c)	52.82	53.9	16.6	53.24	55.7	16.7
PAI (Liu et al., 2024e)	53.49	39.4	11.8	53.27	62.6	18.3
IBD (Zhu et al., 2025a)	52.48	54.1	15.8	54.14	56.4	15.5
VAR (Kang et al., 2025)	52.98	55.1	17.1	-	-	-
AD-HH (Yang et al., 2025)	52.52	44.4	11.5	-	-	-
VAF (Yin et al., 2025)	52.36	55.6	18.1	52.63	55.9	16.9
DeGF (Zhang et al., 2025)	52.72	55.7	16.6	53.06	59.1	17.4
CICD (Zhao et al., 2025)	55.80	45.6	13.1	54.20	45.7	13.2
$\text{SEVI}_{\text{Balanced}}$	53.73	31.9	8.8	53.33	38.0	12.8
$\text{SEVI}_{\text{Focused}}$	**58.00**	**19.3**	**6.1**	**56.89**	**18.3**	**8.2**

Table 2: Results on the COCO subset of DetailCaps. The maximum generation length is 512.

Method	LLaVA-1.5				InstructBLIP			
	CHAIR↓	**Cover↑**	**Hal↓**	**Cog ↓**	**CHAIR↓**	**Cover↑**	**Hal↓**	**Cog ↓**
Regular	11.6	49.7	47.7	4.4	12.4	51.9	52.4	5.0
VCD (Leng et al., 2024)	9.8	51.2	43.8	4.4	9.9	54.0	44.6	4.2
ICD (Wang et al., 2024c)	8.8	51.2	38.7	4.1	9.8	53.9	46.7	5.1
PAI (Liu et al., 2024e)	7.7	49.3	36.9	3.3	11.7	52.8	55.1	5.4
IBD (Zhu et al., 2025a)	9.8	50.5	42.2	4.4	9.0	**56.1**	45.1	4.6
PAI (Liu et al., 2024e)	7.7	49.3	36.9	3.3	11.7	52.8	55.1	5.4
VAR (Kang et al., 2025)	11.7	51.2	48.5	4.8	-	-	-	-
AD-HH (Yang et al., 2025)	9.0	48.0	40.8	3.0	-	-	-	-
VAF (Yin et al., 2025)	11.3	50.2	48.6	4.3	11.5	51.8	50.1	5.1
DeGF (Zhang et al., 2025)	9.1	50.7	39.9	4.1	9.7	54.1	44.5	5.2
CICD (Zhao et al., 2025)	6.6	**52.7**	34.8	2.2	7.1	53.6	35.0	2.3
$\text{SEVI}_{\text{Balanced}}$	**5.6**	48.6	27.6	1.7	**6.0**	51.0	28.5	1.6
$\text{SEVI}_{\text{Focused}}$	6.1	42.3	**20.2**	**0.8**	7.7	42.8	**24.9**	**1.2**

Table 3: Results on AMBER. The maximum generation length is 512.

et al., 2024b) and Qwen2.5-VL-Instruct (Bai et al., 2025)). All models are tested at the 7B parameter scale unless explicitly noted otherwise.

Baselines We conduct a comparison between our method and several SOTA de-hallucination techniques: VCD (Leng et al., 2024), ICD (Wang et al., 2024c), IBD (Zhu et al., 2025a), PAI (Liu et al., 2024e), VAR (Kang et al., 2025), VAF (Yin et al., 2025), AD-HH (Yang et al., 2025), DeGF (Zhang et al., 2025), and CICD (Zhao et al., 2025).

5.2 Main Results

Results on CHAIR CHAIR is a benchmark designed to detect object hallucinations in image captions, relying on human annotations to provide reliable ground truth. Following common practice, we conduct experiments with maximum sequence lengths set to 64 and 512, respectively. The results are shown in Fig. 1. Our method effectively reduces hallucinations by guiding the model's attention toward core information, achieving particularly low hallucination rates under the focused mode. However, we observe that when the model concentrates on critical information, it tends to become more conservative, as reflected by a slight drop in recall. In contrast, the balanced mode allows the model to capture more details while still maintaining a low hallucination rate, resulting in the

Method	LLaVA-Next			Qwen2-VL			Qwen2.5-VL		
	Cs↓	**Ci↓**	**R↑**	**Cs↓**	**Ci↓**	**R↑**	**Cs↓**	**Ci↓**	**R↑**
Regular	32.2	11.0	**56.4**	30.2	8.3	53.1	28.6	9.3	**56.0**
$\text{SEVI}_{\text{Balanced}}$	25.8	9.3	49	20.2	**5.3**	**54.3**	22.2	**7.0**	53.9
$\text{SEVI}_{\text{Focused}}$	**20.8**	**9.9**	37.8	**18.8**	7.0	41.9	**16.4**	7.1	39.2

Table 4: Results CHAIR with more LVLMs. The maximum generation length is 128.

Setting	Cs ↓	Ci ↓	R ↑
PAI with different layers			
None (only CD)	25.4	7.6	61.9
[2-15] (Lower)	20.8	6.8	56.7
[16-31] (Higher)	24.6	8.0	59.3
[2-31] (PAI)	**20.0**	**6.0**	57.6
SEVI combined with PAI			
Regular	24.4	8.9	56.6
$\text{SEVI}_{\text{Focused}}$	14.8	**4.7**	54.1
$\text{SEVI}_{\text{Focused}}$ *w/* PAI	**14.2**	5.3	52.2
$\text{SEVI}_{\text{Balanced}}$	18.8	5.5	59.5
$\text{SEVI}_{\text{Balanced}}$ *w/* PAI	**15.0**	**5.3**	55.4

Table 5: PAI increase visual attention from the 2nd layer to the final layer ([2-31]). *None (only CD)* denotes the use of contrastive decoding alone, without increasing visual attention.

Setting	LLaVA-1.5		InstructBLIP	
	Cs↓	**Ci↓**	**Cs↓**	**Ci↓**
Regular	54.6	16.4	62.6	19.5
w/o CICD	38.6	13.9	44.4	20.3
w/o Core Heads	39.0	9.7	45.0	13.3
w/o Global Heads	22.2	7.0	23.2	11.6
w/o Two-Stage Opt.	18.4	9.6	35.2	10.7
$\text{SEVI}_{\text{Focused}}$	**17.8**	**5.5**	**18.8**	**8.4**

Table 6: Ablation study on CHAIR with *Focused* mode. *w/o Two-Stage Opt.* represents applying core semantic heads to non-semantic heads. We find that this approach disrupts the model's internal representations and leads to response collapse. To improve the smoothness of attention alignment, we set $\omega = 1$ in this setting.

best overall performance. In addition, our method yields the most pronounced improvement on the 13B model relative to the 7B model, indicating its effectiveness in harnessing the latent capacity of larger models for visual understanding.

Results on DetailCaps DetailCaps evaluates the correctness of captions in terms of objects, attributes, and relationships, incorporating both exact-match and soft-match metrics. The experimental results are presented in Tab. 2. Our method outperforms existing approaches in both hallucination rate and overall caption correctness. Moreover, the focused mode achieves a higher CPAURE score, indicating not only a lower incidence of hallucinations but also greater accuracy in the described content. This highlights the practical value of our approach for real-world applications.

Results on AMBER AMBER carefully selects high-quality images to construct its benchmark and provides a more fine-grained evaluation of object hallucinations. The experimental results are shown in Tab. 3. Our method achieves a significantly lower hallucination rate while maintaining a comparable Cover score to other approaches, resulting in the best overall performance.

5.3 Visual or Semantic Representations, Which Should LVLMs Focus On?

Some works (Liu et al., 2024e; Zhu et al., 2025a) emphasize the importance of visual attention, whereas we encourage the model to focus more on semantic representations. To investigate whether the model should prioritize visual or semantic representations, we take PAI (Liu et al., 2024e) as a contrast, which directly increases the model's attention to visual representations.

As we analyze in Fig. 1 and Fig. 2, visual information resides in visual representations during the early layers of LVLMs and is gradually integrated into semantic representations in later layers. Therefore, we hypothesize that **LVLMs should focus on visual representations in the lower layers and shift to semantic representations in the higher layers**. We perform ablation studies on both lower and higher layers within the PAI to better understand the layer-wise impact. The results at the top

of Tab. 5 are consistent with our hypothesis, indicating that it is more beneficial for the model to focus on visual representations in the lower layers. Furthermore, we explore combining PAI with our method by applying PAI to the lower layers (layers [2–15]) to enhance visual attention, while adopting our approach in the higher layers (layers [16–31]). The results at the bottom of Tab. 5 further demonstrate our hypothesis.

5.4 Effectiveness on more LVLMs

We evaluated our method on several state-of-the-art LVLMs, with the experimental results presented in Tab. 4. Across these models, our method continues to demonstrate strong hallucination suppression capabilities, with both modes exhibiting their expected effectiveness. These results validate the excellent generalization ability of our approach and highlight its practical potential as a training-free solution.

5.5 Ablation Study

Through a two-stage optimization paradigm, we reallocate the model's attention toward the semantic representations, encouraging it to focus on the core visual information and thereby reducing hallucinations. In addition, we employ CICD to address the issue of language priors that aggravate when the model over-focuses on semantic representations. We conduct ablation studies on these three components, with the results presented in Tab. 6. The core semantic heads play a pivotal role in guiding the model's focus, and removing them in the ablation study leads to a noticeable increase in hallucination rate. The global semantic heads help modulate the cross-modal attention distribution, aligning it with the flow of visual information, which also contributes significantly to hallucination reduction. The combination of both components has a synergistic effect, and incorporating the CICD method to mitigate language priors, amplified by overattending to semantic representations, further enhances the performance of the model. Additionally, the two-stage optimization paradigm also has a positive effect, making the optimization process more stable and smooth. The experimental results confirm the effectiveness and soundness of our method's design, highlighting the value of each component.

6 Conclusion

We systematically analyze the information flow in LVLMs and verify that visual information is indeed integrated into the semantic representations. However, the model's attention remains predominantly focused on the visual representation. This inconsistency impairs the model's visual understanding and contributes to hallucination. To address this, we propose Semantic-Enhanced Visual Interpretation (SEVI), a method that guides the model's attention toward the core components of semantic representations through a two-stage optimization process. Extensive experiments demonstrate that our approach significantly mitigates hallucinations.

While our method optimizes the attention distribution in LVLMs based on the underlying information flow, it does not directly enhance the flow mechanism itself. In future work, we plan to explore more efficient strategies for visual information propagation.

Limitations

Our method guides the model to focus on the most critical information, thereby reducing the occurrence of hallucinations. During our experiments, we observed a trade-off mechanism in the model's focusing process: excessive focus leads to a more conservative image captioning. In an effort to avoid hallucinated content, the model may overlook some details, which is reflected in a slight decrease in recall. Although our method demonstrates superior overall performance—showing a significant advantage on comprehensive metrics such as CAPTURE—it still faces a trade-off between reducing hallucinations and generating more details. To address this issue, we introduce hyperparameters to control the model's level of conservativeness, allowing users to manually adjust the behavior based on specific application scenarios. Furthermore, as a training-free approach, our method offers greater usability but is inherently limited by the performance ceiling of the model itself.

Acknowledgments

We thank all reviewers for their detailed reviews and valuable comments. This work is supported by Zhongguancun Academy Project No.20240103.

References

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Proc. of ACL*, pages 4190–4197.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou,

and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Cong Chen, Mingyu Liu, Chenchen Jing, Yizhou Zhou, Fengyun Rao, Hao Chen, Bo Zhang, and Chunhua Shen. 2025. Perturbollava: Reducing multimodal hallucinations with perturbative visual training. In *Proc. of ICLR*.

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. In *Proc. of NIPS*, pages 49250–49267.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *Proc. of AAAI*, 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 *Proc. of ACL*.

Hongyu Hu, Jiyuan Zhang, Minyi Zhao, and Zhenbang Sun. 2023. Ciem: Contrastive instruction evaluation method for better instruction tuning. In *NIPS Workshop on Instruction Tuning and Instruction Following*.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. 2025. See what you are told: Visual attention sink in large multimodal models. In *Proc. of ICLR*.

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 *Proc. of CVPR*, pages 13872–13882.

Jiaming Li, Jiacheng Zhang, Zequn Jie, Lin Ma, and Guanbin Li. 2025. Mitigating hallucination for large vision language model by inter-modality correlation calibration decoding. *arXiv preprint arXiv:2501.01926*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proc. of EMNLP*, pages 292–305.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Proc. of ECCV*, pages 740–755.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024a. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *Proc. of ICLR*.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024c. Improved baselines with visual instruction tuning. In *Proc. of CVPR*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024d. Llava-next: Improved reasoning, ocr, and world knowledge.

Shi Liu, Kecheng Zheng, and Wei Chen. 2024e. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *Proc. of ECCV*, pages 125–140.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, and 1 others. 2024. Gpt-4o cystem card.

Zhibo Ren, Huizhen Wang, Muhua Zhu, Yichao Wang, Tong Xiao, and Jingbo Zhu. 2023. Overcoming language priors with counterfactual inference for visual question answering. In *Proc. of CCL*, pages 600–610.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proc. of EMNLP*, pages 4035–4045.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, and 1 others. 2024. Aligning large multimodal models with factually augmented rlhf. In *Findings of ACL*, pages 13088–13110.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proc. of NIPS*.

Bin Wang, Fan Wu, Xiao Han, Jiahui Peng, Huaping Zhong, Pan Zhang, Xiaoyi Dong, Weijia Li, Wei Li, Jiaqi Wang, and 1 others. 2024a. Vigc: Visual instruction generation and correction. In *Proc. of AAAI*, pages 5309–5317.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and 1 others. 2023a. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, and 1 others. 2023b. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023c. Label words are anchors: An information flow perspective for understanding in-context learning. In *Proc. of EMNLP*, pages 9840–9855.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024c. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of ACL*, pages 15840–15853.

Yike Wu, Yu Zhao, Shiwan Zhao, Ying Zhang, Xiaojie Yuan, Guoqing Zhao, and Ning Jiang. 2022. Overcoming language priors in visual question answering via distinguishing superficially similar instances. In *Proc. of COLING*, pages 5721–5729.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. In *Proc. of ICLR*.

Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. 2025. Understanding and mitigating hallucination in large vision-language models via modular attribution and intervention. In *Proc. of ICLR*.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision).

Qinghao Ye, Xianhan Zeng, Fu Li, Chunyuan Li, and Haoqi Fan. 2025. Painting with words: Elevating detailed image captioning with benchmark and alignment learning. In *Proc. of ICLR*.

Hao Yin, Guangzong Si, and Zilei Wang. 2025. Clearsight: visual signal enhancement for object hallucination mitigation in multimodal large language models. In *Proc. of CVPR*, pages 14625–14634.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2024. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105.

Bohan Zhai, Shijia Yang, Xiangchen Zhao, Chenfeng Xu, Sheng Shen, Dongdi Zhao, Kurt Keutzer, Manling Li, Tan Yan, and Xiangjun Fan. 2024. Halle-switch: Rethinking and controlling object existence hallucinations in large vision-language models for detailed caption.

Ce Zhang, Zifu Wan, Zhehan Kan, Martin Q Ma, Simon Stepputtis, Deva Ramanan, Russ Salakhutdinov, Louis-Philippe Morency, Katia Sycara, and Yaqi Xie. 2025. Self-correcting decoding with generative feedback for mitigating hallucinations in large vision-language models. In *Proc. of ICLR*.

Jianfei Zhao, Feng Zhang, Xin Sun, and Chong Feng. 2025. Cross-image contrastive decoding: Precise, lossless suppression of language priors in large vision-language models. *arXiv preprint arXiv:2505.10634*.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024. Analyzing and mitigating object hallucination in large vision-language models. In *Proc. of ICLR*.

Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. 2025a. Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding. In *Proc. of CVPR*, pages 1624–1633.

Younan Zhu, Linwei Tao, Minjing Dong, and Chang Xu. 2025b. Mitigating object hallucinations in large vision-language models via attention calibration. *arXiv preprint arXiv:2502.01969*.

A Two-Stage Optimization

We present the algorithm of our method in Algorithm 1.

B Detailed Experiments

B.1 Benchmarks

We evaluate the performance of our method across three widely adopted multimodal hallucination benchmarks on the image captioning task. These include CHAIR, DetailCap, and AMBER.

CHAIR (Rohrbach et al., 2018) evaluates the proportion of hallucinated objects—those generated by the model but not present in the reference annotations. Following prior work, we randomly sample 500 images from the MSCOCO (Lin et al., 2014) dataset for evaluation. CHAIRs and CHAIRi are the main metrics to evaluate hallucination:

$$\begin{aligned} \text{CHAIRs} &= \frac{|\text{Hallucinated Objects}|}{|\text{All Objects}|}, \\ \text{CHAIRi} &= \frac{|\text{Hallucinated Sentences}|}{|\text{All Sentences}|} \end{aligned} \tag{8}$$

DetailCaps (Ye et al., 2025) is a fine-grained image captioning benchmark, accompanied by ground-truth detail captions generated by GPT-4V (Yang et al., 2023), Gemini 1.5 Pro (Team et al., 2024), and GPT-4o (OpenAI et al., 2024) for evaluation purposes. It comprises 4,870 images from various datasets; we use a subset of 700 images from MSCOCO in our experiments. CAPTURE evaluates the alignment between generated and reference captions by computing F1 scores using both hard and soft matching across three semantic aspects: entities ($F1_{\text{obj}}$), attributes ($F1_{\text{attr}}$), and relations ($F1_{\text{rel}}$). The final score is calculated as a weighted average:

$$\text{CAPTURE} = \frac{\alpha F1_{\text{obj}} + \beta F1_{\text{attr}} + \gamma F1_{\text{rel}}}{\alpha + \beta + \gamma} \tag{9}$$

where $\alpha = 5$, $\beta = 5$, and $\gamma = 2$.

AMBER (Wang et al., 2023a) contains 1,004 carefully curated high-quality images, each with manually annotated objects. AMBER contains multiple evaluation metrics: *CHAIR*, *Cover*, *Hal*, and *Cog*. Given a list of annotated objects $A_{obj} = obj_1^A, obj_2^A, \cdots, obj_n^A$ and a set of generated objects R'_{obj}, each metric is defined as follows:

$$\begin{aligned} \text{CHAIR} &= 1 - \frac{\text{len}(R'_{obj} \cap A_{obj})}{\text{len}(R'_{obj})}, \\ \text{Cover} &= \frac{\text{len}(R'_{obj} \cap A_{obj})}{\text{len}(A_{obj})}, \\ \text{Hal} &= \frac{|\text{CHAIR} > 0|}{|\text{All Captions}|}, \\ \text{Cog} &= \frac{\text{len}(R'_{obj} \cap H_{obj})}{\text{len}(R'_{obj})} \end{aligned} \tag{10}$$

where H_{obj} denotes the set of hallucinated target objects generated by LVLMs, and All Captions refers to the total number of generated captions.

B.2 Baselines

We conduct a comparison between our method and several SOTA de-hallucination techniques:

- **VCD** (Leng et al., 2024) introduces Gaussian noise into images to activate language priors, thereby constructing negative contexts and removing these priors through contrastive decoding. However, this approach leads to a loss of visual information. Moreover, the use of noisy images deviates from the model's training distribution, potentially causing performance degradation.

- **ICD** (Wang et al., 2024c) constructs negative contexts by designing adversarial instructions and applies contrastive decoding to mitigate their influence. Like VCD, it faces challenges such as visual information loss and performance bias.

- **IBD** (Zhu et al., 2025a) strengthens the model's focus on visual information by using the original context as a negative reference, and further refines contrastive decoding via inter-layer and inter-context consistency mechanisms.

- **VAR** (Kang et al., 2025) reallocates attention from sink visual tokens to other visual tokens, allowing the model to capture more detailed visual information.

- **VAF** (Yin et al., 2025) rebalances the attention allocation between instructions and visual inputs, redirecting attention from the textual instructions toward the visual information.

Setting	Cs↓	Ci↓	R ↑	R-Cs-Ci↑
$\kappa = 0.15$	21.6	7.1	57.6	28.9
$\kappa = 0.2$	17.8	5.5	56.9	**33.6**
$\kappa = 0.25$	12.6	6.3	49.1	30.2
$\kappa = 0.3$	11.0	5.4	44.8	28.4

Table 7: Explore on hyperparameter κ with the *Focused* mode.

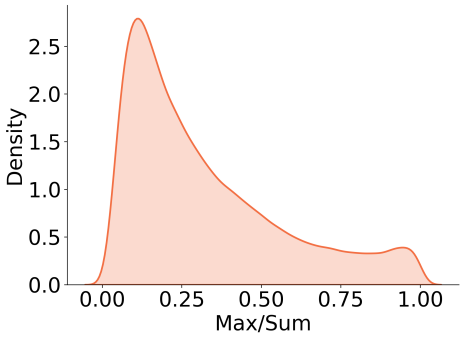

Figure 5: Semantic attention peaks. We plot the KDE (Kernel Density Estimation) of the peak attention weights for attention heads focusing on semantic representations. The x-axis represents the proportion of the highest-attended representation's attention weight among all semantic representations, while the y-axis denotes the density.

- **DeGF** (Zhang et al., 2025) generates negative contexts via cross-modal back-translation and strengthens consistency signals throughout the contrastive decoding process.
- **CICD** (Zhao et al., 2025) leverages the consistency of language priors across images by using different images to construct negative contexts. It detects detrimental priors via consistency analysis and removes them through contrastive decoding, while retaining beneficial priors essential for accurate understanding.

B.3 Hyperparameters

We investigate the appropriate setting of hyperparameters using LLaVA-1.5-7B. The parameter κ serves as the threshold for distinguishing between core semantic heads and global semantic heads. We analyze the peak attention values of semantic heads, as shown in Fig. 5. Based on this analysis, we explore the impact of κ under the focused mode, with the results presented in Tab.7. A higher threshold encourages the model to attend to more central information, resulting in a lower hallucination rate but also a reduced recall. To balance these trade-offs, we define a heuristic metric to evaluate the model's overall performance. Based on the experimental results, we set $\kappa = 0.2$.

The attention adjustment strength ω and the starting layer (SL) for attention optimization are interdependent. Therefore, we perform a joint search over these two hyperparameters, with the results summarized in Tab.8. Based on two heuristic evaluation metrics, we select two representative hyperparameter configurations, corresponding to the Focused mode (ω=4, SL=5) and Balanced modes (ω=0.5, SL=9).

C Case Studies

To demonstrate the effectiveness of our method in mitigating hallucinations, we provide qualitative case studies. We select one simple image (Fig. 6) and one complex image (Fig. 7) as case studies. The captions were generated by LLaVA-1.5, with hallucinated content highlighted in red and image-consistent content highlighted in green.

It can be observed that Regular Decoding produces a large amount of hallucinated content, even in the simple image. In contrast, our method effectively reduces the occurrence of such hallucinations. Both the Focused and Balanced modes significantly mitigate hallucinations; however, the Focused mode makes the model overly conservative, potentially overlooking fine-grained details. For example, in Fig. 7, the Focused mode fails to capture items such as *knife* and *fork*, whereas the Balanced mode not only reduces hallucinations but also preserves more detailed information.

Algorithm 1 Attention Distribution Alignment within a Layer

Input: attention weights $\mathbf{W} \in \mathbb{R}^{H \times L \times L}$,
the end indexes of visual tokens e,
hyperparameters κ and ω
Output: aligned attention weights $\hat{\mathbf{W}}$

$\mathbf{W}_{\text{cur}} \leftarrow \mathbf{W}[:, -1, :]$ # Attention weights of the last token
$\mathbf{W}_{\text{sf}} \leftarrow \mathrm{Softmax}(\mathbf{W}_{\text{cur}})$ # Apply softmax normalization
$\mathbf{W}_S \leftarrow \mathbf{W}_{\text{sf}}[:, e+1:]$ # Attention to semantic representations
$\mathbf{W}_O \leftarrow \mathbf{W}_{\text{sf}}[:, :e+1]$ # Attention to other representations
Head categorization:
$H_S \leftarrow (\sum \mathbf{W}_S > \sum \mathbf{W}_O)$ # Semantic heads
$H_O \leftarrow \neg H_S$ # Other heads
$H_c \leftarrow (\max \mathbf{W}_S > \kappa \cdot \sum \mathbf{W}_S)$ # Heads focusing on core information
$H_g \leftarrow \neg H_c$ # Heads fail to focus on core information
$H_{S_g} \leftarrow H_S \cap H_g$ # Global semantic heads
$H_{S_c} \leftarrow H_S \cap H_c$ # Core semantic heads
# Two-Stage Optimization:
if $|H_O| > 0$ **and** $|H_{S_g}| > 0$ **then**
  # Guide the other heads using the global semantic heads
  $m_1 \leftarrow \frac{1}{|H_{S_g}|} \sum \mathbf{W}_{\text{cur}}[H_{S_g}, :]$ # Average pooling
  $\mathbf{W}_{\text{cur}}[H_O, :] \leftarrow (\mathbf{W}_{\text{cur}}[H_O, :] + \omega \cdot m_1)/(1+\omega)$
end if
if $|H_{S_c}| > 0$ **and** $|H_{S_g}| > 0$ **then**
  # Guide the global semantic heads using the core semantic heads
  $m_2 \leftarrow \max \mathbf{W}_{\text{cur}}[H_{S_c}, :]$ # Max pooling
  $\mathbf{W}_{\text{cur}}[H_{S_g}, :] \leftarrow (\mathbf{W}_{\text{cur}}[H_{S_g}, :] + \omega \cdot m_2)/(1+\omega)$
end if
$\mathbf{W}[:, -1, :] \leftarrow \mathbf{W}_{\text{cur}}$ # Update attention weights
Return: $\mathbf{W}$

Setting	CHAIRs↓	CHAIRi↓	Recall ↑	R-Cs-Ci↑	2R-Cs-Ci↑
ω=0.5, SL=2	32.8	8.7	68.0	26.5	94.5
ω=0.5, SL=5	34.4	9.4	68.8	25.0	93.8
ω=0.5, SL=9	31.4	9.1	67.8	27.3	**95.1**
ω=0.5, SL=17	45.2	12.3	74.3	16.8	91.1
ω=1, SL=2	29.4	7.7	64.4	27.3	91.7
ω=1, SL=5	27.4	8.0	64.7	29.3	94.0
ω=1, SL=9	26.0	6.7	63.7	31.0	94.7
ω=1, SL=17	47.0	13.5	74.3	13.8	88.1
ω=2, SL=2	23.4	7.3	60.5	29.8	90.3
ω=2, SL=5	23.6	6.7	59.9	29.6	89.5
ω=2, SL=9	23.6	7.1	61.2	30.5	91.7
ω=2, SL=17	47.8	13	74.9	14.1	89.0
ω=4, SL=2	18.0	6.1	55.5	31.4	86.9
ω=4, SL=5	17.8	5.5	56.9	**33.6**	90.5
ω=4, SL=9	21.6	7.1	57.6	28.9	86.5
ω=4, SL=17	47.0	13.0	74.1	14.1	88.2

Table 8: Results of grid search. *SL* stands for *Start Layer*, indicating the layer from which attention optimization begins. *Cs*, *Ci*, and *R* separately represent *CHAIRs*, *CHAIRi*, and *Recall*.

Regular

The image features a red fire hydrant sitting in the middle of a field with its base elevated up in the sand. The fire hydrant is surrounded by plants and likely serves its purpose in case of emergencies. It is situated near a concrete wall which appears to be a part of the surrounding landscaping. The hydrant is prominently visible in the scene, with no other objects or potted plants providing a sense of perspective within the area.

Balanced mode

The image features a red fire hydrant prominently sitting in a rock bed or rock garden. The hydrant is situated on a gravel bed and is surrounded by green plants and shrubs. The vibrant red fire hydrant stands out against the lush greenery, creating a harmonious blend of colors and textures. The combination of the vivid red hue and the lush plant life creates a serene and peaceful atmosphere.

Focused mode

The image features a red fire hydrant sitting in a rock bed or gravel in front of some plants and bushes. The fire hydrant is red in color, and its bright red hue contrasts sharply with the surrounding plants, adding a beautiful contrast to the garden scene. The bright red color of the fire hydrant creates a striking contrast with its surroundings, making it a striking addition to the scene.

Figure 6: Case study I.

Regular

The image shows a middle-aged couple seated together at a dining table, enjoying a meal. In front of them, several bowls are filled with food, and wine glasses are positioned around the table. Other tableware items include a fork, knife, and spoons, placed near the people. The couple appears to be sitting in a casual restaurant setting. Additionally, there is a chair on the right side of the table for another person, although no one else is visible in the scene. A person can be seen in the background, possibly a waiter or a fellow patron at the restaurant.

Balanced mode

The image depicts an older couple sitting together at a dining table, enjoying a meal of soup and salad. They have multiple plates, cups, and bowls filled with food in front of them. Various utensils such as forks, spoons, and knives can be seen placed on the table, ready for use. Additional tableware is present, including a vase, and a bottle can be spotted on the table. The couple seems to be enjoying each other's company and sharing the dining experience.

Focused mode

The image shows two elderly people, a man and a woman, sitting at a dining table and enjoying a meal together. They appear to be smiling while posing for the camera. On the table, there are various bowls containing different types of food, along with multiple spoons placed in and around them. Additionally, there are several wine glasses on the table, indicating a variety of wines may have been served with the meal. Overall, it's a joyful and happy moment.

Figure 7: Case study II.