

CAI: Caption-Sensitive Attention Intervention for Mitigating Object Hallucination in Large Vision-Language Models

Anonymous ACL submission

Abstract

Although Large Vision-Language Models (LVLMs) have demonstrated powerful capabilities in interpreting visual information, they frequently produce content that deviates from visual information, leading to object hallucination. To tackle this, recent works mostly depend on expensive manual annotations and training cost, or significantly increase inference time. In this work, we observe that LVLMs' attention to visual information is significantly stronger when answering caption queries compared to non-caption queries. Inspired by this phenomenon, we propose Caption-sensitive Attention Intervention (CAI), a training-free, plug-and-play hallucination mitigation method that leverages the attention activation pattern in response to caption queries to enhance LVLMs' visual perception capability. Extensive experimental results across four benchmarks covering both discriminative and generative tasks, demonstrate that CAI achieves state-of-the-art (SOTA) hallucination mitigating performance only with minimal additional inference cost.

1 Introduction

Despite the continuous advancements in the performance of large vision-language models (LVLMs) in recent years, it is widely observed that LVLMs frequently generate content that conflicts with the corresponding visual information, leading to hallucination (Sahoo et al., 2024; Huang et al., 2023).

Previous works (Arif et al., 2025; Bi et al., 2024a) show that LVLMs' insufficient attention to visual information is considered a key factor in hallucination. To tackle this, recent works for mitigating hallucination mostly use contrastive decoding strategy (Leng et al., 2024; Zhong et al., 2024) which arises high inference latencies, or train LVLMs using carefully designed data (You et al., 2023; Yu et al., 2024a) which incurs expensive manual annotation and computation cost. To address the aforementioned limitations, we focus on

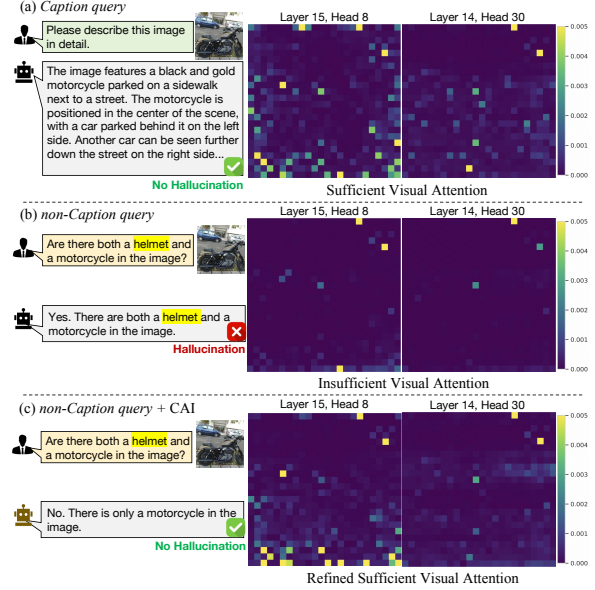


Figure 1: The visualization of attention weights at image patch level across different conversations. LVLM correctly generates the detailed content of the image in response to the caption query, but exhibits hallucination (e.g., "helmet") when answering the non-caption query. CAI refines LVLM's visual attention patterns from insufficient to sufficient, effectively enhancing visual perception capability and mitigating hallucination.

exploring how to enhance LVLMs' perception capability by providing sufficient attention to visual information. In this work, as shown in Figure 1 (a) and (b), we reveal a critical phenomenon: visual attention across particular attention heads was significantly enhanced when fed caption compared to non-caption queries. We term these attention heads as caption-sensitive attention heads. As an enhancement of their visual attention leads to a corresponding improvement in LVLM's perception capability, we believe that these heads are also visually sensitive. Therefore, we manage to probe and refine these caption-sensitive attention heads.

Inspired by the aforementioned phenomenon, we propose Caption-sensitive Attention Intervention

(CAI), a training-free, plug-and-play method, to refine caption-sensitive attention heads outputs during inference to enhance LVLMs’ fine-grained visual perception capability and mitigate hallucination. First, we identify the optimal caption query from candidates, which activates the model’s inherent visual perception capability with the minimal necessary attention weight shift cost. Secondly, following previous work (Li et al., 2024), we train binary classifiers to identify caption-sensitive attention heads and compute their attention output heads shifts, which quantify the differences from non-caption to caption queries and serve as a vision-centric optimization direction. Finally, we apply the precomputed attention output shifts to intervene caption-sensitive attention heads during inference. As shown in Figure 1 (b) and (c), after using CAI, LVLM enhances sufficient visual attention and effectively mitigates hallucination.

We evaluate the performance of CAI across multiple discriminative and generative tasks, using models such as LLaVA-1.5-7b (Liu et al., 2024a), Qwen-VL-Chat (Bai et al., 2023), and LLaVA-NeXT (Liu et al., 2024b). On the POPE (Li et al., 2023) benchmark, the accuracy and the F1 score improve by 5.14% and 5.50% on average. On the MME (Fu et al., 2023) hallucination subset, the scores increase by 64.3 points on average. Furthermore, hallucination rates decrease by 7.8% on the MMHalBench (Sun et al., 2023), while the informativeness of the generated responses improves.

The main contributions can be summarized as:

- (1) Our work is the first to explicitly reveal the impact of caption versus non-caption queries on the attention activation patterns of LVLMs.
- (2) We propose CAI, a training-free, plug-and-play method significantly mitigates object hallucination in LVLMs by refining caption-sensitive attention head outputs during the inference.
- (3) Comprehensive experimental results demonstrate that CAI effectively mitigates hallucination.

2 Quantitative Analysis of the Effect of Caption Queries on Visual Attention

To better validate the motivation of our CAI method that caption queries can help LVLMs refine visual attention activation patterns, we constructed a quantitative analysis experiment. We sample 1,000 images from the MS-COCO dataset (Lin et al., 2014). For each image, we propose one caption query and two different non-caption queries (non-caption-1 &

non-caption-2) to analyze differences attributable to query types. We compute the *Change Rate* to quantify differences in visual attention weight changes. The specific calculation and other experimental details can be found in the Appendix G.

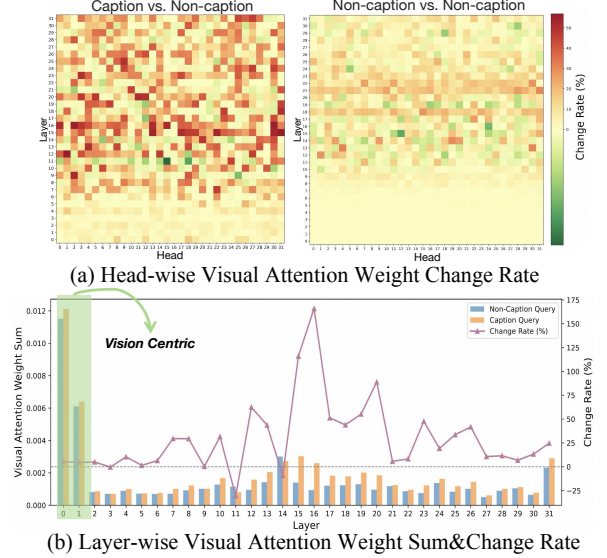


Figure 2: A systematic quantitative analysis from head-wise (a) and layer-wise (b) on visual attention weights. The comparison shows that the caption query significantly enhanced the visual attention of LLaVA-1.5-7b.

As illustrated in Figure 2, caption queries have a more significant impact on the LVLM’s visual attention weights compared with non-caption queries. 65.92% of the attention heads and 30 of the 32 layers exhibit an enhancement in visual attention weight. As shown in Figure 2 (b), the first two layers allocate significantly more visual attention compared to other layers, suggesting that they are centered on visual information and relatively insensitive to input queries. In deeper layers, LVLM’s enhanced visual attention is an important and non-negligible reason for stronger perception capability.

3 Methods

3.1 Task Formulation

We consider a LVLM parametrized by θ . The model receives as input a visual input $V = \{v_1, v_2, \dots, v_m\}$ and a textual query $T = \{t_1, t_2, \dots, t_n\}$, where m and n denote the sequence lengths of the visual input and textual inputs. The textual and visual inputs are concatenated together to form the first layer input $H^1 = \text{concat}(V, T) \in \mathbb{R}^{(m+n) \times d}$ for the L layers $\times H$ heads language decoder.

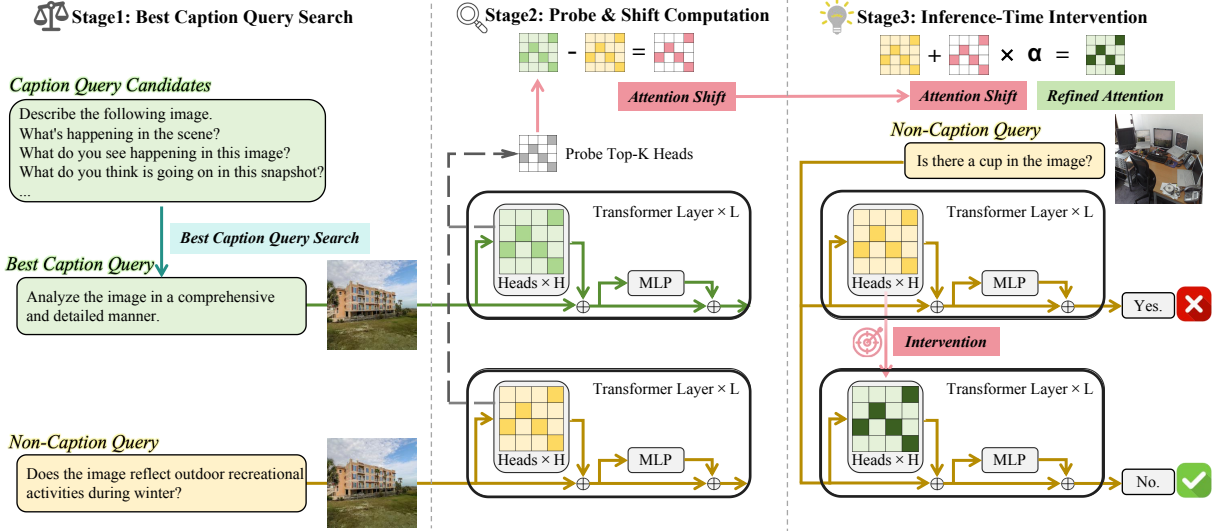


Figure 3: An overview of the CAI method. Each square in the matrix represents the attention head output. Squares with dark green color indicate refined attention head outputs. CAI consists of three stages: (1) §3.2 Best caption query search algorithm is designed to seek the best optimization target query with minimal necessary attention weight shift. (2) §3.3 The original and modified attention outputs are used to identify caption-sensitive attention heads and compute attention output shift vectors. (3) §3.4 Precomputed attention shift vectors are applied to the top K caption-sensitive attention heads during inference, thereby enhancing their visual attention and activating the model’s inherent fine-grained visual perception to mitigate hallucination.

During the forward pass, the input \mathbf{H}^l received by the h -th attention head at l -th layer is linearly transformed using independent weight matrices to generate the Query, Key and Value matrices, denoted as $\mathbf{Q}_{(l,h)} \in \mathbb{R}^{(m+n) \times d}$, $\mathbf{K}_{(l,h)} \in \mathbb{R}^{(m+n) \times d}$ and $\mathbf{V}_{(l,h)} \in \mathbb{R}^{(m+n) \times d}$, where d denotes the head hidden dimensions. The generated Query, Key, and Value matrices are then used to compute the attention score, attention weight matrix and attention output as follows:

$$\dot{\mathbf{A}}_{(l,h)} = \frac{\mathbf{Q}_{(l,h)} \mathbf{K}_{(l,h)}^T}{\sqrt{d}}, \mathbf{A}_{(l,h)} = \text{softmax}(\dot{\mathbf{A}}_{(l,h)}),$$

$$\mathbf{O}_{(l,h)} = \mathbf{A}_{(l,h)} \mathbf{V}_{(l,h)}. \quad (1)$$

At each layer, the updated hidden state \mathbf{H}^{l+1} is then computed by adding the residual connection to the output of the multi-head attention mechanism:

$$\mathbf{H}^{l+1} = \mathbf{H}^l + \sum_{h=1}^H \mathbf{O}_{(l,h)} \cdot \mathbf{W}_o^l, \quad (2)$$

where $\mathbf{W}_o^l \in \mathbb{R}^{Hd \times d}$ is the learnable weight matrix for the linear transformation applied after concatenating the outputs from all H attention heads. Finally, the model predicts the next token in an autoregressive manner based on the last layer output.

3.2 Best Caption Query Search Algorithm

This module aims to seek the best caption query, which induces the minimal necessary attention weight shift to activate the LVLMs’ fine-grained visual perception capabilities. For a single VQA question, we separately use a certain caption query T from J candidate queries and a non-caption query T' paired with same image V as inputs during the forward pass to compute caption-sensitive attention weight matrix $\mathbf{A}_{(l,h)}$ and non-caption attention weight matrix $\mathbf{A}'_{(l,h)}$. The attention weight shift matrix can be computed as:

$$\mathbf{A}_{shift} = \sum_{l=1}^L \sum_{h=1}^H (\mathbf{A}_{(l,h)} - \mathbf{A}'_{(l,h)}). \quad (3)$$

For a VQA dataset with a batch size of B and a caption query candidate list with length of J , the index of the best query j in the candidate list can then be calculated as:

$$\arg \min_j \sum_{b=1}^B \mathbf{A}_{shift}^{b,j} \quad s.t. \quad j \in J, \quad (4)$$

where $\mathbf{A}_{shift}^{b,j}$ denotes the attention weight shift matrix when answering the b -th VQA using the j -th caption query.

3.3 Caption-Sensitive Attention Heads Probe

This module aims to identify caption-sensitive attention heads, which exhibit significant differences in attention outputs when responding to caption and non-caption queries. We focus on the LVLMS' attention output shift of visual information, aiming to minimize the influence of textual semantic information during the probing process. To achieve this, we set the last token's text-related attention scores of each attention head to $-\infty$ during the forward pass, and compute the modified attention output:

$$\dot{\mathbf{A}}_{(l,h)}[m : m + n] = -\infty, \quad (5)$$

$$\hat{\mathbf{O}}_{(l,h)} = \text{softmax}(\dot{\mathbf{A}}_{(l,h)})\mathbf{V}_{(l,h)}, \quad (6)$$

$$\tilde{\mathbf{O}}_{(l,h)} = \hat{\mathbf{O}}_{(l,h)}[m + n]. \quad (7)$$

For a dataset with a batchsize of B , the last token's modified attention output of b' -th VQA problem when answering the caption query and non-caption query are denoted as $\tilde{\mathbf{O}}_{(l,h)}^b$ and $\tilde{\mathbf{O}}_{(l,h)}^{b'}$. Respectively, the last token's origin attention output are denoted as $\mathbf{O}_{(l,h)}^b$ and $\mathbf{O}_{(l,h)}^{b'}$.

For each head, we use the B pairs of modified attention output as input to train a binary classifier $f_{l,h}(\cdot)$ that predicts whether the input sentence is a caption query:

$$\arg \min_{f_{l,h}(\cdot)} \sum_{b=1}^B \mathcal{L}(f_{l,h}(x_b), y_b), \quad (8)$$

where $x_b \in \{\tilde{\mathbf{O}}_{(l,h)}^b, \tilde{\mathbf{O}}_{(l,h)}^{b'}\}$ denotes the input of classifier and $y_b \in \{0, 1\}$ denotes the category of query. We then select the top K binary classifiers with the highest accuracy. In order to compute the optimization direction for each head, the attention output shift vector is computed as follows:

$$\mathbf{S}_{(l,h)} = \frac{1}{B} \sum_{b=1}^B (\mathbf{O}_{(l,h)}^b - \mathbf{O}_{(l,h)}^{b'}). \quad (9)$$

3.4 Intervention at Inference Time

This module aims to refine the top K attention heads output that are most sensitive to caption queries at inference-time. We leverage the precomputed shift vectors to refine these heads from insufficient visual attention states to sufficient states, thereby enhancing the model's fine-grained visual perception capability and mitigate hallucination.

At each layer, the updated hidden state after intervention is computed as:

$$\mathbf{H}^{l+1} = \mathbf{H}^l + \sum_{h=1}^H (\mathbf{O}_{(l,h)} + \mathbb{I}_{(l,h)} \alpha \mathbf{S}_{(l,h)}) \cdot \mathbf{W}_o^l, \quad (10)$$

where $\mathbb{I}_{(l,h)}$ is a gating function, assigning a value of 1 to attention heads with top k highest accuracy, and 0 to the others. α represents the intensity of the intervention.

4 Experiments

4.1 Experimental Setup

We comprehensively evaluate the methods for both discriminative and generative tasks to measure the effectiveness and robustness of the methods.

Discriminative Tasks:

POPE employs a binary question-answering format, inquiring LVLMS to answer if a special object exists in the given image. Following previous works, we adopt Accuracy and F1 score as the evaluation metrics.

MME serves as a comprehensive tool for assessing the capabilities of LVLMS across both 10 perception tasks and 4 cognition tasks. Consequently, task scores are reported as the evaluation metrics.

Generative Tasks:

CHAIR (Rohrbach et al., 2018) is a widely used metric for assessing object hallucination in responses of LVLMS. Following previous work, We use the MS-COCO Chair subset with the prompt "Please describe this image in detail." to evaluate the hallucination mitigating capabilities of LVLMS. The CHAIR metric comprises two important indicators, denoted as CHAIR_i and CHAIR_s , with the following calculation formulas:

$$\text{CHAIR}_i = \frac{|\{\text{Hallucinationed objects}\}|}{|\{\text{All objects mentioned}\}|}$$

$$\text{CHAIR}_s = \frac{|\{\text{Sentence with hallucination objects}\}|}{|\{\text{All sentence}\}|}$$

MMHal-Bench comprises 96 meticulously designed VQA questions, which evaluates response-level hallucination rate and informativeness. It asks GPT-4 to compare model outputs with human responses and object labels for evaluation.

Baselines. We adopt the widely used LLaVA-1.5-7b, Qwen-VL-Chat and LLaVA-NeXT (Liu et al., 2024b) as our baseline LVLMS. We compared CAI with the following SOTA training-free methods:

Setting	Method	LLaVA-1.5-7b		Qwen-VL-Chat		LLaVA-NeXT	
		Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score
Random	Regular	83.29	81.33	84.63	82.61	84.78	86.43
	VCD	87.73	87.16	86.93	85.46	88.76	89.57
	OPERA	89.20	88.81	85.71	84.64	90.27	89.71
	PAI	86.33	84.56	85.38	85.54	88.40	87.16
	VTI	89.50	88.89	86.73	85.59	89.23	88.68
	CAI(ours)	89.87(+6.58)	89.43(+8.10)	88.17(+3.54)	87.31(+4.70)	90.68(+5.90)	90.42(+3.99)
Popular	Regular	81.88	80.06	83.63	81.53	83.23	84.77
	VCD	85.38	85.06	85.17	83.68	87.01	87.70
	OPERA	86.64	86.62	84.82	83.99	87.16	87.68
	PAI	85.33	83.62	84.20	83.10	86.65	86.99
	VTI	87.36	86.69	85.67	84.48	87.33	87.16
	CAI(ours)	88.32(+6.44)	87.95(+7.89)	87.73(+4.10)	86.84(+5.31)	89.53(+6.30)	89.24(+4.47)
Adversarial	Regular	78.96	77.57	81.03	79.30	81.19	82.50
	VCD	80.88	81.33	83.10	82.04	84.80	85.23
	OPERA	81.24	81.38	82.67	79.89	85.20	85.54
	PAI	83.17	81.67	82.19	82.06	84.32	83.68
	VTI	82.57	82.11	83.13	82.16	85.35	84.52
	CAI(ours)	84.27(+5.31)	84.41(+6.84)	84.33(+3.30)	83.92(+4.62)	85.97(+4.78)	86.07(+3.57)

Table 1: Main results on POPE tasks. We evaluate the accuracy and F1 Score of various LVLMs on the MS-COCO POPE tasks. The best performances within each setting are bolded. **CAI(ours)** demonstrates the best hallucination mitigation performance among several methods.

Method	LLaVA-1.5-7b					Qwen-VL-Chat					LLaVA-NeXT				
	Exist.	Count	Pos.	Color	Total	Exist.	Count	Pos.	Color	Total	Exist.	Count	Pos.	Color	Total
Regular	175.7	124.7	114.0	151.0	565.4	170.0	135.0	123.3	170.0	598.3	180.0	105.0	150.0	151.7	586.7
VCD	180.3	131.7	125.0	155.0	592.0	180.0	133.3	131.7	175.0	620.0	185.0	125.0	133.3	168.3	611.6
OPERA	165.0	116.0	133.3	149.0	563.3	180.0	140.0	138.3	175.0	633.3	183.8	121.3	155.0	162.1	622.2
PAI	190.0	148.3	126.7	160.0	625.0	175.0	141.6	132.5	177.5	626.6	185.0	128.3	148.3	170.8	632.4
VTI	185.0	140.0	135.0	165.7	619.0	180.0	142.5	133.0	178.0	633.5	186.7	126.7	150.0	172.5	635.9
CAI(ours)	190.0	141.6	140.0	170.0	641.6	185.0	150.0	133.3	180.0	648.3	190.0	133.3	155.0	175.0	653.3

Table 2: Main results on the hallucination subset of MME. The best performances within each setting are bolded.

(1) **Baselines tailored for decoding:** VCD (Leng et al., 2024) contrasts model logits derived from original and distorted visual input to reduce the over-reliance on statistical bias and unimodal priors. OPERA (Huang et al., 2024) introduces a penalty term on the model logits.

(2) **Baselines utilizing inference-time intervention:** PAI (Liu et al., 2024d) intervenes on attention heads by leveraging their original direction and optimizes the output distribution during decoding to mitigate language bias. VTI (Liu et al., 2024c) mitigates hallucination by steering layer-level latent space representations during inference to enhance visual feature stability.

Despite prior findings (Bi et al., 2024b) indicating the significant role of attention heads in visual perception, there is a lack of approaches that analyze at head level and do not rely on specific decoding strategies (which increase inference time). To address these limitations, our CAI probes and refines attention heads with minimal additional inference cost by exploiting the differential attention

activation pattern between caption and non-caption queries, thereby achieving superior results.

Implementation Details. In our experiments, we utilized 13 caption queries and 100 VQA from LLaVA pretrain dataset to search the best caption. Then we utilized 1000 VQA from LLaVA pretraining dataset pairs with the searched best caption query and non-caption queries to identify caption-sensitive attention heads and computed the attention shift vectors. For each attention head, SVM (Cortes, 1995) was used as the classifier and 2-fold cross-validation was performed to evaluate its accuracy. More detailed experimental procedures are provided in Appendix A.

4.2 Main Results

Comprehensive evaluations demonstrate that our method exhibits superior hallucination mitigation capabilities in discriminative and generative tasks.

Result on POPE. Table 1 presents the POPE tasks results under three different experimental settings.

(1) **SOTA performance.** Compared with other

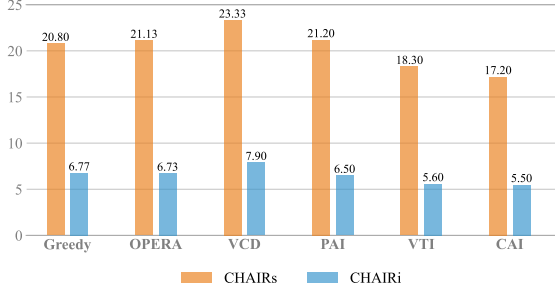


Figure 4: Main result of LLaVA-1.5-7b on MS-COCO CHAIR task. Smaller values of CHAIR_i and CHAIR_s indicate that the method demonstrates stronger hallucination mitigation capabilities at instance and sentence levels. Max_new_tokens is set to be 64.

methods, CAI achieved superior hallucination mitigation performance across all experimental configurations. Specifically, it leads to an average improvement of 6.11% in accuracy and 7.61% in F1 score for LLaVA-1.5-7b, 3.65% and 4.88% for Qwen-VL-Chat, 5.66% and 4.01% for LLaVA-NeXT, resulting in SOTA hallucination mitigation effects. (2) **Easy to deploy in open-source LVLMS.** During the best caption query searching, caption-sensitive attention heads probe and shift vector computation stages, CAI selected images, caption query candidates and non-caption queries from LLaVA-1.5-7b pre-training dataset, which are outside the domain of the benchmark datasets. Despite this cross-domain discrepancy, our method achieved significant improvements even when applied to Qwen-VL-Chat and LLaVA-NeXT, indicating that CAI does not rely on specific models or data and deploy in open-source LVLMS easily.

Results on MME. Table 2 and Figure 9 respectively present the experimental results for the MME hallucination subset and full set. **Our method effectively mitigates hallucination while preserving the LVLMS’ other foundational capabilities.** On the MME hallucination subset, our method achieved the best improvements across all capabilities with score increases of 76.2 for LLaVA-1.5-7b, 50.0 for Qwen-VL-Chat and 66.6 for LLaVA-NeXT. On the full MME dataset, performance improved on 13 out of 14 perception and reasoning tasks, with an overall score increase of 197.63 for LLaVA-1.5-7b.

Results on CHAIR. Figure 4 demonstrates that our method significantly reduces both sentence-level and instance-level hallucination in responses to caption queries. Specifically, we observed re-

Method	LLaVA-1.5-7b		Qwen-VL-Chat	
	Score \uparrow	VH Rate% \downarrow	Score \uparrow	VH Rate% \downarrow
Greedy	1.86	63.5	2.93	41.1
VCD	2.12	54.2	2.77	39.2
OPERA	2.15	54.2	2.94	38.4
PAI	2.27	53.2	2.87	39.5
VTI	2.33	52.2	2.99	38.4
CAI(ours)	2.43	51.0	3.04	38.0

Table 3: Main result on MMHal-Bench.

ductions of 3.6% in the CHAIR_s metric and 1.27% in the CHAIR_i metric. By employing the CAI, we precisely identify the attention heads that play a crucial role in visual perception under the caption task and accurately estimate their optimization directions. Although these attention heads have already been activated and demonstrate the ability to perceive visual information, applying CAI method can further strengthen visual attention, enhancing LVLMS’ perceptual capabilities and resulting in better performance on the caption task.

Results on MMHal-Bench. Table 3 presents our method effectively reduces the hallucination rate in responses to non-caption queries while enhancing informativeness, outperforming several inference-time intervention methodes.

5 Analysis and Discussions

5.1 Inference Latency

Method	TTFT(ms)	TPOT(ms)	Acc(%)
LLaVA-1.5-7b	99.8 1.0 \times	36.0 1.0 \times	78.96
+VCD	160.1 1.6 \times	96.8 2.7 \times	80.88
+OPERA	109.8 1.1 \times	69.5 1.9 \times	81.24
+VDGD	377.8 3.8 \times	340.9 9.5 \times	65.82
+CAI(ours)	102.2 1.0 \times	36.5 1.0 \times	84.50

Table 4: Inference latency (Time to First Token, Time Per Output Token) and the accuracy on MS-COCO adversarial POPE of different methods.

As shown in Table 4, although VDGD (Ghosh et al., 2024) attempts to mitigate hallucination using captioning capability, directly using caption description significantly increases computational cost and forces the model to process longer context, leading to a performance drop on the POPE task. Compared to contrastive decoding-based methods which trade-off speed for accuracy, CAI implicitly utilizes captioning capabilities, achieving faster inference speed and better performance.

5.2 Necessity of the Search Algorithm

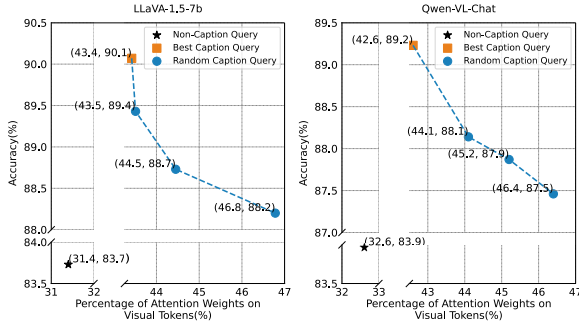


Figure 5: The accuracies of baselines and CAI with different caption queries on GQA Random POPE task.

To better understand the necessity of the search algorithm, we will focus on analyzing the following two issues:

One potential question is, *Why does CAI prefer using a single caption query instead of combining multiple queries?* The answer lies in the fact that different caption queries activate different paths during inference. Combining multiple caption queries causes interference between the activated paths, preventing the performance improvements that could be achieved by using any single caption query.

Another key question is, *Why does CAI select the query with minimal attention weight shift as the best?* The primary goal of CAI is to refine the outputs of caption-sensitive attention heads without significantly altering LVLMs’ existing behaviors. By minimizing the attention weight shift, CAI strikes a balance between enhancing visual perception and maintaining the integrity of other foundational capabilities. As illustrated in Figure 5, the experimental results demonstrate that LVLMs achieve the best performance using CAI search strategy. Certain caption query without careful selection may lead LVLMs to excessively focus on visual information, preventing it from achieving maximum performance improvement.

5.3 Distribution of Probed Attention Heads

As illustrated in Figure 6, we visualize the classification accuracies across 32×32 attention heads. We observe that caption-sensitive attention heads are concentrated primarily between the 7th and 20th layers, which provides corroborating evidence for the quantization experiments detailed in Figure 2. These layers are critical for balancing visual perception and semantic understanding within the

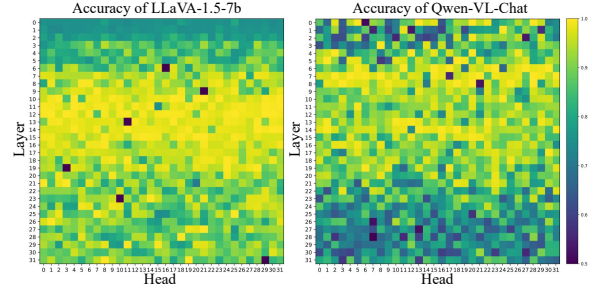


Figure 6: The accuracies of classifiers.

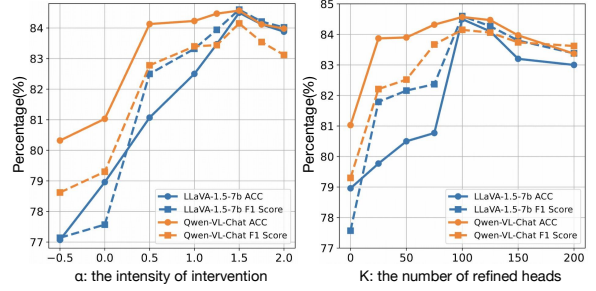


Figure 7: Ablation study of α and K .

model. By refining the output of these attention heads, CAI significantly enhances LVLMs’ visual perception capability and mitigate hallucination.

5.4 Implications of Hyperparameters

CAI method primarily relies on two key hyperparameters: the intensity of intervention α and the number of refined attention heads K . We performed a series of ablation experiments using greedy decoding on the MS-COCO Adversarial POPE dataset. As shown in Figure 7, the key implications can be summarized as follows:

(1) Impact of α : A negative value reduces the model’s attention to visual tokens, which in turn diminishes its performance in hallucination mitigation. When α is small, the attention intervention is insufficient, resulting in only marginal improvements in model performance. A large α leads to insufficient attention to textual information, resulting in a decline in performance.

(2) Impact of K : Applying intervention to too few attention heads fails to effectively influence the activation pathways of visual information, resulting in limited improvements in hallucination mitigation. Conversely, intervening in too many attention heads can disrupt critical attention activation paths that are unrelated to visual perception and play essential roles in other foundational capabilities, leading to performance degradation.

See Appendix E for impact of over-intervention.

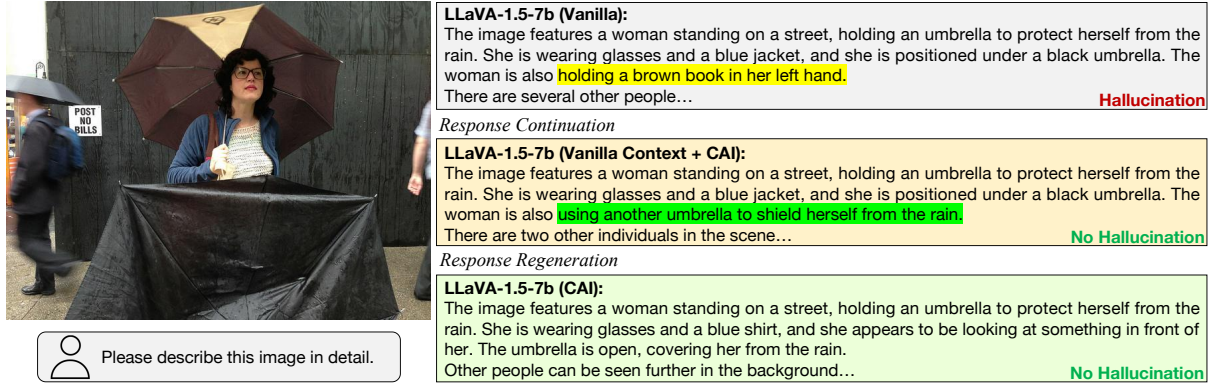


Figure 8: LLaVA-1.5-7b demonstrated hallucination when addressing caption query of MS-COCO CHAIR dataset. Both response continuation and regeneration for the same query provided by CAI effectively mitigate hallucination.

5.5 Case Study on CHAIR

CAI proves effective in mitigating hallucination even when addressing caption queries. CAI strengthens the visual information attention of attention heads responsible for captioning tasks, further enhancing LVLMs’ captioning capability. As shown in Figure 8, CAI effectively mitigates hallucination through sentence continuation and regeneration. See the Appendix H for more cases.

6 Related Works

6.1 Large Vision-Language Models

Several powerful LVLMs based on open-source LLM backbones combined with visual encoders have achieved impressive capabilities through extensive vision-language pretraining. Furthermore, recent searches have further improved model performance by employing high-resolution visual encoders (Hong et al., 2024) and exploring reinforcement learning methods, such as RLHF (Yu et al., 2024a). Closed-source models, such as GPT-4o (Hurst et al., 2024) and Gemini 1.5 (Reid et al., 2024) have demonstrated even more powerful performance. However, recent LVLMs still suffer from hallucination problems. Addressing how to cost-effectively mitigate hallucination is still a critical problem that requires further exploration.

6.2 Mitigating Hallucination in LVLMs

Current methods for mitigating hallucination in LVLMs can be broadly categorized into two types: data-driven training methods and training-free methods. Training-based methods typically involve introducing novel training objectives (Chen et al., 2024a) and utilizing carefully curated datasets (Gunjal et al., 2024; Liu et al., 2023; Yu et al.,

2024b; You et al., 2023). For training-free methods, the main strategies include designing decoding techniques (Leng et al., 2024; Chen et al., 2024b; Chuang et al., 2023; Huang et al., 2024; Zhong et al., 2024) during the inference phase, leveraging language or visual prompts (Lee et al., 2023; An et al., 2024), incorporating external tools or knowledge sources (Zhao et al., 2024) and correct the generation (Yin et al., 2024). Furthermore, a key approach to reduce hallucination is to address attention deficits by adjusting or using decoding strategies (An et al., 2024; Gong et al., 2024; Xing et al., 2024). PAI (Liu et al., 2024d) intervenes in attention heads by leveraging the direction and magnitude of their original outputs, and optimizes the output distribution during decoding to mitigate hallucinations. VTI (Liu et al., 2024c) reduces hallucinations by steering layer-level latent space representations during inference to enhance the stability of vision features. However, our work is the first to explicitly reveal the impact of caption vs. non-caption queries on the attention activation patterns of LVLMs and mitigate hallucination by applying caption-sensitive attention intervention at head level during the inference.

7 Conclusion

In this paper, we presented CAI, a training-free method that refines caption-sensitive attention heads outputs for non-caption queries towards outputs for caption queries during the inference, thereby leveraging LVLMs’ inherent fine-grained visual perception capabilities to mitigate object hallucination. CAI demonstrates strong generalizability and can be applied to several open-source LVLMs. Consistent performance improvements across diverse benchmarks highlight its robustness.

8 Limitations

While CAI demonstrates significant effectiveness in mitigating object hallucinations in LVLMs, several limitations should be acknowledged to provide a balanced perspective on its applicability and scope. First, CAI relies on the availability of suitable caption queries to identify the optimal attention shift. Expanding the caption query candidate list could address this issue. Moreover, LVLMs that do not rely on multi-head attention mechanisms, or those employing non-standard visual-textual alignment strategies, may not benefit from CAI’s intervention method. We will address the above issues in future work.

References

Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Guang Dai, Ping Chen, and Shijian Lu. 2024. Agla: Mitigating object hallucinations in large vision-language models with assembly of global and local attention. *arXiv preprint arXiv:2406.12718*.

Kazi Hasan Ibn Arif, Sajib Acharjee Dip, Khizar Husain, Lang Zhang, and Chris Thomas. 2025. Fixing imbalanced attention to mitigate in-context hallucination of large vision-language model. *arXiv preprint arXiv:2501.12206*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Jing Bi, Junjia Guo, Yunlong Tang, Lianggong Wen, Zhang Liu, and Chenliang Xu. 2024a. [Unveiling visual perception in language models: An attention head analysis approach](#). *ArXiv*, abs/2412.18108.

Jing Bi, Junjia Guo, Yunlong Tang, Lianggong Bruce Wen, Zhang Liu, and Chenliang Xu. 2024b. Unveiling visual perception in language models: An attention head analysis approach. *arXiv preprint arXiv:2412.18108*.

Beitao Chen, Xinyu Lyu, Lianli Gao, Jingkuan Song, and Heng Tao Shen. 2024a. Alleviating hallucinations in large vision-language models through hallucination-induced optimization. *arXiv preprint arXiv:2405.15356*.

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

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola:

Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*.

Corinna Cortes. 1995. Support-vector networks. *Machine Learning*.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. 2024. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14303–14312.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. [Mme: A comprehensive evaluation benchmark for multimodal large language models](#). *ArXiv*, abs/2306.13394.

Sreyan Ghosh, Chandra Kiran Reddy Evuru, Sonal Kumar, Utkarsh Tyagi, Oriol Nieto, Zeyu Jin, and Dinesh Manocha. 2024. [Vdgd: Mitigating lvlm hallucinations in cognitive prompts by bridging the visual perception gap](#). *Preprint*, arXiv:2405.15683.

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

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.

Wenyi Hong, Weihang Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. 2024. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

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.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

- Zhehan Kan, Ce Zhang, Zihan Liao, Yapeng Tian, Wenming Yang, Junyuan Xiao, Xu Li, Dongmei Jiang, Yaowei Wang, and Qingmin Liao. 2024. Catch: Complementary adaptive token-level contrastive decoding to mitigate hallucinations in lvlms. *arXiv preprint arXiv:2411.12713*.
- Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10.
- Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. 2023. Volcano: mitigating multimodal hallucination through self-feedback guided revision. *arXiv preprint arXiv:2311.07362*.
- 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.
- 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*.
- Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36.
- 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*.
- 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 *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.
- Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. 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. 2024b. *Llava-next: Improved reasoning, ocr, and world knowledge*.
- Sheng Liu, Haotian Ye, and James Zou. 2024c. Reducing hallucinations in vision-language models via latent space steering. *arXiv preprint arXiv:2410.15778*.
- Shi Liu, Kecheng Zheng, and Wei Chen. 2024d. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. *arXiv preprint arXiv:2407.21771*.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2024e. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.
- Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.
- Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.
- Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. A comprehensive survey of hallucination in large language, image, video and audio foundation models. *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11709–11724.
- Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.
- Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. *arXiv preprint arXiv:2403.18715*.
- Yun Xing, Yiheng Li, Ivan Laptev, and Shijian Lu. 2024. Mitigating object hallucination via concentric causal attention. *arXiv preprint arXiv:2410.15926*.
- 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.
- Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. 2023. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*.
- Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwan He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2024a. Rlhf-v: Towards

trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816.

Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. 2024b. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*.

Linxi Zhao, Yihe Deng, Weitong Zhang, and Quanquan Gu. 2024. Mitigating object hallucination in large vision-language models via classifier-free guidance. *arXiv preprint arXiv:2402.08680*.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. *arXiv preprint arXiv:2210.07197*.

Weihong Zhong, Xiaocheng Feng, Liang Zhao, Qiming Li, Lei Huang, Yuxuan Gu, Weitao Ma, Yuan Xu, and Bing Qin. 2024. [Investigating and mitigating the multimodal hallucination snowballing in large vision-language models](#). *Preprint*, arXiv:2407.00569.

Guanyu Zhou, Yibo Yan, Xin Zou, Kun Wang, Aiwei Liu, and Xuming Hu. 2024. Mitigating modality prior-induced hallucinations in multimodal large language models via deciphering attention causality. *arXiv preprint arXiv:2410.04780*.

Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. 2024. *Ibd*: Alleviating hallucinations in large vision-language models via image-biased decoding. *arXiv preprint arXiv:2402.18476*.

A Additional Experimental Details

All datasets used in this paper are licensed under a [Creative Commons Attribution 4.0 License](#).

A.1 Data Source

Although our method does not rely on specific data, we separately specify the sources of the data used in the experiments for the sake of reproducibility.

A.1.1 Data of Best Query Search

In the best caption search algorithm, we use the top 100 VQA samples from the complex reasoning data in the LLaVA-1.5-7b pre-training dataset. From this, we obtain non-caption queries and their corresponding images. Additionally, we maintain a list of 13 candidate caption queries, some of which are manually generated and others are derived from the pre-trained instructions of LLaVA-1.5-7b. The caption query candidates are listed as follows:

"What do you see happening in this image?", "What do you think is going on in this snapshot?", "Can you elaborate on the elements of the picture provided?", "Describe the following image.", "What's happening in the scene?", "Analyze the image in a comprehensive and detailed manner.", "Write a detailed description of the given image.", "What is this photo about?", "Explain the visual content of the image in great detail.", "What are the key elements in this picture?", "Can you describe the main features of this image for me?", "Please describe this image in detail.", "Generate the caption in English:"

In the experiments, the best caption query for LLaVA-1.5-7b is *"Analyze the image in a comprehensive and detailed manner."* and the best caption query for Qwen-VL-Chat is *"Please describe this image in detail."*

A.1.2 Data of Probe and Shift Computation

We extracted the first 1,000 samples from the complex reasoning data in the LLaVA-1.5-7b pre-training dataset. The questions from these samples were treated as non-caption queries.

A.2 Detailed Experimental Setup

In the experiment of POPE, 'regular' refers to the direct sampling setting. We used direct sampling decoding and set $\alpha = 1.5$ and $K = 100$ in the main experiments.

B Comparison with More Advanced Methods

We selected LLaVA-1.5-7b as the baseline model and compared CAI with more advanced models including VCD (Leng et al., 2024), ICD (Wang et al., 2024), OPERA (Huang et al., 2024), Woodpecker (Yin et al., 2024), M3ID (Favero et al., 2024), DAMRO (Gong et al., 2024), IMCCD (Li et al., 2025), CATCH (Kan et al., 2024), IBD (Zhu et al., 2024) and CAUSALMM (Zhou et al., 2024). The results of CAI compared with several state-of-the-art methods on MS-COCO Adversarial POPE are shown in Table 7.

C Detailed Experimental Results of MME

Detailed experimental results on MME perception and cognition can be found in Table 8 and Table 9.

D Detailed Experimental Results of Ablation

Detailed results of the ablation experiments can be found in Table 5 and Table 6.

α	LLaVA-1.5-7b		Qwen-VL-Chat	
	Accuracy	F1-Score	Accuracy	F1-Score
-0.50	77.07	77.14	80.32	78.62
0.00	78.96	77.57	81.03	79.30
0.50	81.07	82.50	84.13	82.78
1.00	82.50	83.32	84.23	83.40
1.25	83.47	83.94	84.47	83.44
1.50	84.50	84.60	84.57	84.15
1.75	84.90	84.21	84.12	83.54
2.00	85.10	84.02	83.98	83.12

Table 5: Detailed results of α on MS-COCO Adversarial POPE dataset.

K	LLaVA-1.5-7b		Qwen-VL-Chat	
	Accuracy	F1-Score	Accuracy	F1-Score
0	78.96	77.57	81.03	79.30
25	79.77	81.79	83.87	82.21
50	80.50	82.16	83.90	82.52
75	80.77	82.37	84.32	83.67
100	84.50	84.60	84.57	84.15
125	84.10	84.27	84.47	84.06
150	83.20	83.80	83.97	83.74
200	83.00	83.38	83.37	83.62

Table 6: Detailed results of K on MS-COCO Adversarial POPE dataset.

Method	Random		Popular		Adversarial		Average	
	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score
Regular	83.29	81.33	81.88	80.06	78.96	77.57	81.38	79.65
VCD (CVPR'24)	87.73	87.16	85.38	85.06	80.88	81.33	84.66	84.52
ICD (EMNLP'24 findings)	89.56	89.68	86.16	86.76	79.71	81.70	85.14	86.05
OPERA (CVPR'24)	89.20	88.81	86.64	86.62	81.24	81.38	85.70	85.60
Woodpecker (SCIS'24)	87.67	86.45	80.67	79.72	80.67	80.00	83.00	82.05
M3ID (CVPR'24)	86.20	84.51	84.77	83.17	82.53	81.14	84.50	82.94
DAMRO (EMNLP'24)	88.20	87.29	85.67	84.98	82.07	81.90	85.31	84.72
IMCCD (arXiv'25)	89.23	88.68	86.73	86.13	82.87	82.77	86.27	85.86
CATCH (ECCV'24)	90.43	90.13	87.07	86.56	83.17	83.18	86.89	86.62
VDD (arXiv'24)	90.00	88.79	85.91	84.40	83.52	82.20	86.48	85.13
CAUSALMM (ICLR'25)	88.93	88.10	87.13	87.26	83.70	82.78	86.59	86.05
CAI(ours)	89.87	89.43	88.32	87.95	84.27	84.41	87.49	87.22

Table 7: Result compared with more advanced methods on MS-COCO POPE.

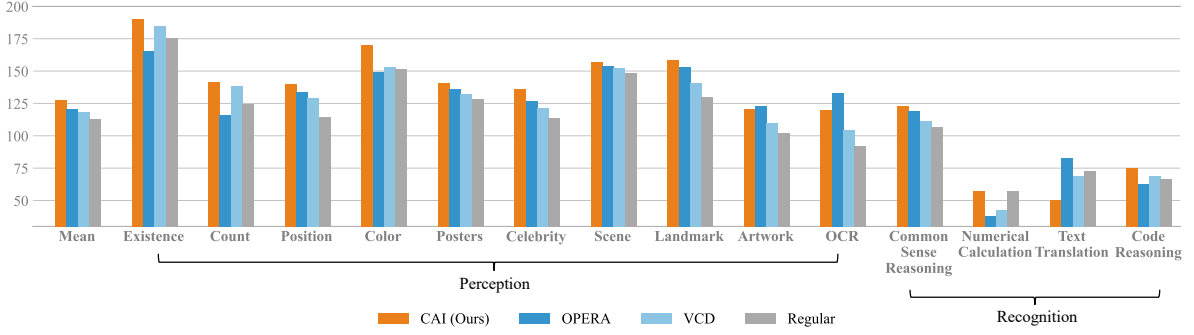


Figure 9: Main results on the MME full set. CAI leads to the best enhancement in LVLMs’ perception capacities while preserving their recognition competencies.

Method	Artwork	Celebrity	Color	Count	Existence	Landmark	OCR	Position	Posters	Scene	Total
Regular	102.20	113.59	151.00	124.67	175.67	129.95	92.00	114.00	127.82	148.30	1279.20
VCD	109.60	120.94	153.00	138.33	184.66	140.45	104.00	128.67	132.11	152.20	1363.96
OPERA	122.50	126.76	149.00	116.00	165.00	152.75	132.50	133.33	136.05	154.00	1387.89
CAI(ours)	120.25	135.88	170.00	141.67	190.00	158.50	120.00	140.00	140.48	157.00	1473.78

Table 8: Results on all MME perception-related tasks. The best performance of each setting is **bolded**.

Method	Coding Reasoning	Commonsense Reasoning	Numerical Calculation	Text Translation	Total
Regular	66.38	106.43	57.00	72.50	302.31
VCD	68.50	111.29	42.64	68.50	290.93
OPERA	62.50	119.29	37.50	82.50	301.79
CAI(ours)	75.00	122.86	57.50	50.00	305.36

Table 9: Results on all MME recognition-related tasks. The best performance of each setting is **bolded**.

Method	CHAIRs	CHAIRi	PPL	Coherence	Fluency
Greedy	20.80	6.77	3.97	0.998500	0.805269
CAI	17.20	5.50	4.11	0.998352	0.791763
CAI (over-intervention)	18.60	6.00	4.23	0.998180	0.809675

Table 10: Impact of over-intervention on CHAIR

E Impact of Over-intervention

In this work, we have provided a detailed discussion of the probe for intervention heads, the number of intervention heads, and the intervention strength. Experimental results were used to determine various hyper-parameters to avoid over-intervention. Following prior work, we employ UniEval (Zhong et al., 2022) and perplexity (PPL) computation to assess the coherence and fluency. As the experimental results show in Table 10, whether applying the CAI or the CAI (over-intervention) with excessive intervention (hyper-parameters set to 2 times as the normal), the PPL, Coherence and Fluency scores remains stable without significant fluctuations. This indicates that the CAI method does not sacrifice the model’s semantic coherence and contextual fluency.

F Domain Generalization Performance

Domain	Dataset	Method	Accuracy
Medical	VQA-RAD	Greedy	54.18%
		CAI	58.17%
OCR	MMBench	Greedy	74.31%
		CAI	77.54%

Table 11: Results on VQA-RAD and MMBench OCR subset.

In domain-specific tasks, the CAI method demonstrates certain generalization ability to some extent. Although caption queries are general instructions, they are extensively used during model pretraining. Activating the relevant attention patterns facilitates fine-grained visual information capture, thereby enhancing downstream task performance. To evaluate CAI’s effectiveness in specific domains, we selected VQA-RAD (Lau et al., 2018) from the medical domain and the MMBench (Liu et al., 2024e) OCR subset. The experimental results of LLaVA-1.5-7b, as presented in the table 11, show consistent improvements over the baseline, indicating the CAI method’s generalization ability.

G Detailed Experimental Setup of Quantitative Analysis

We sample 1,000 images from the MS-COCO dataset (Lin et al., 2014). For each image, we propose one caption query and two different non-caption queries (non-caption-1 & non-caption-2) to analyze differences attributable to query types.

We consider a LVLM parametrized by θ . The model receives as input a textual query $\mathbf{T} = \{t_1, t_2, \dots, t_n\}$ and a visual input $\mathbf{V} = \{v_1, v_2, \dots, v_m\}$, where n and m denote the sequence lengths of the text and visual inputs. The text and vision inputs are concatenated together to form the first layer input $\mathbf{H}^1 = \text{concat}(\mathbf{V}, \mathbf{T}) \in \mathbb{R}^{(m+n) \times d}$ for the L layers $\times H$ heads decoder. For an image, the last input token’s visual attention weight of H -th head in L -th layer $\text{Sum}_{(l,h)}$ can be computed as:

$$\mathbf{A}_{(l,h)} = \text{softmax}\left(\frac{\mathbf{Q}_{(l,h)} \mathbf{K}_{(l,h)}^T}{\sqrt{d}}\right), \quad (11)$$

$$\text{Sum}_{(l,h)} = \sum_{i=1}^m \mathbf{A}_{(l,h)}^{-1}[i], \quad (12)$$

where the $\mathbf{Q}_{(l,h)}$ and $\mathbf{K}_{(l,h)}$ are the Query and Key matrixes of the k -th head in l -th layer, $\mathbf{A}_{(l,h)}^{-1}[i]$ is the last input token’s attention weight of the i -th input token. For a dataset of B samples, the sum of visual attention weight can be computed as:

$$\mathbf{S}_{(l,h)} = \sum_{b=1}^B \text{Sum}_{(l,h)}. \quad (13)$$

Then we record the sum of visual attention weights from the last input token for three types of queries: $S_{(l,h)}^{\text{cap}}$ for caption query, $S_{(l,h)}^{\text{non-1}}$ for non-caption query 1 and $S_{(l,h)}^{\text{non-2}}$ for non-caption query 2. The head-wise Change Rate $\text{Rate}_{(l,h)}$ and layer-wise Change Rate $\text{Rate}_{(l)}$ can be computed as:

$$\text{Rate}_{(l,h)} = \frac{S_{(l,h)}^{\text{cap}} - S_{(l,h)}^{\text{non-1}}}{S_{(l,h)}^{\text{non-1}}}, \quad (14)$$

$$\text{Rate}_{(l)} = \frac{\sum_{h=1}^H (S_{(l,h)}^{\text{cap}} - S_{(l,h)}^{\text{non-1}})}{\sum_{h=1}^H S_{(l,h)}^{\text{non-1}}}. \quad (15)$$

By comparison, we find that visual attention across particular attention heads was significantly enhanced when fed caption compared to non-caption queries. These results provide strong support for our proposed motivation.

H Additional Case Studies

More case studies are shown as follows.



LLaVA-1.5-7b (Vanilla):

The image depicts a rainy day in a city, with a wet street and sidewalk. The street is lined with trees, and there are several cars parked along the side of the road. A bus is also visible on the street, adding to the urban atmosphere. There are several people walking on the sidewalk, some of them carrying umbrellas to shield themselves from the rain. A traffic light is present in the scene, and a fire hydrant can be seen further down the street. The street appears to be a popular spot for people to walk, as there are several benches placed along the sidewalk. One of the benches is located near the center of the scene, while another is situated closer to the right side.

GPT-4 Evaluation:

The LMM's response is detailed and generally accurate in describing the street, the trees, the parked cars, the people with umbrellas, and the wet conditions suggestive of a recent rainfall. However, the model hallucinates when it infers that there is a bus present in the picture, layers additional detail by saying that there are benches along the sidewalk, and specifies the location of a fire hydrant. These elements are not mentioned in the human-generated description or implied by the image contents provided.

Rating: 1, somewhat informative, with hallucination.



Describe the image in detail.

Golden Answer:

The picture depicts a city street, recently dampened by rain, with a sidewalk lined by trees. There are several cars parked on the right, while others move along the road. A few pedestrians can be seen at a distance on the sidewalk, some of them holding umbrellas to protect themselves from the rain. The trees offer shade and lend a touch of nature to the urban setting. The scene exudes a sense of calm and serenity, emphasized by the glistening wet street and the nearly vacant walkway.

LLaVA-1.5-7b (CAI):

The image depicts a rainy day in a city, with a wet street and sidewalk. The street is lined with trees, and there are several cars parked along the side of the road. A bus is also visible on the street, adding to the urban atmosphere.

There are multiple traffic lights in the scene, with one near the center of the image and others scattered around the area. A fire hydrant can be seen on the sidewalk, and a stop sign is located further down the street.

A few people are walking along the sidewalk, with one person closer to the center of the image and two others further down the street. The overall scene captures the feeling of a rainy day in a city.

GPT-4 Evaluation:

The LMM's response is accurate and does not hallucinate any information. It captures the main details described in the human-generated response, such as the city street, parked cars, pedestrians, and rainy atmosphere. It also mentions other details not mentioned in the human response, like the bus, the traffic lights, a fire hydrant, and a stop sign. These additional details add to the descriptiveness of the LMM's response. However, it does not capture the sense of "calm and serenity" mentioned in the human-generated response, which may reduce its informativeness slightly.

Rating: 5, very informative, no hallucination.

Figure 10: Caption query case of LLaVA-1.5-7b on MMHal-Bench.



LLaVA-1.5-7b (Vanilla):

The image features a zoo enclosure with a group of zebras and giraffes standing together. There are three zebras in the foreground, with one of them being a baby zebra. Behind them, there are two giraffes, one of which is a baby giraffe. The animals are standing near a rock wall, and there is a large rock in the background.

The enclosure is surrounded by a fence, and there is a small building in the background. The zebras and giraffes appear to be enjoying their time together in the zoo.

GPT-4 Evaluation:

The LMM's response is somewhat informative as it describes the presence of zebras and giraffes near a rock wall, mentioning a small building and a fence surrounding the enclosure, all of which aligns with the image's content. However, the response also has some minor inaccuracies, such as the claim of baby zebras and giraffes being present, and stating that there are only two giraffes when there are actually four. Additionally, there is a slight hallucination in that it claims the animals are "enjoying their time", which is an anthropomorphic attribution that is not directly observable from the image. Therefore, the response is rated as somewhat informative, with minor hallucinations.

Rating: 1, somewhat informative, with hallucination.



Describe the image in detail.

Golden Answer:

The image depicts three zebras and four giraffes on a sandy terrain, with trees in the background and a stone structure to the right. Surrounding the scene is fencing, suggesting a controlled environment like a zoo or nature reserve.

LLaVA-1.5-7b (CAI):

The image features a zoo enclosure with a group of zebras and giraffes standing together. There are three zebras in the scene, with one positioned closer to the left side, another in the middle, and the third one on the right side. A giraffe is standing near the left side of the enclosure, while another giraffe is positioned more towards the center.

The enclosure is surrounded by a rock wall, and there is a large rock in the middle of the scene. The animals appear to be enjoying their time together, possibly in a shaded area.

GPT-4 Evaluation:

The LMM's response provides a detailed description of the image, identifying the animals present (zebras and giraffes) and giving a sense of their positioning within the enclosure. However, it neglects to specify the number of giraffes, which were included in the standard human-generated answer. It also makes a claim of the animals possibly enjoying their time together in a shaded area, an assumption not explicitly made in the image content. Nonetheless, it captures the main elements in the image with no significant hallucination.

Rating: 4, somewhat informative, no hallucination.

Figure 11: Another caption query case of LLaVA-1.5-7b on MMHal-Bench.



How was the weather when this picture was captured?

Golden Answer:

When this photo was taken, it was dark at night. This photo shows a horse race at night.

LLaVA-1.5-7b (Vanilla):

The weather in the picture was dark and cloudy.

GPT-4 Evaluation:

The LMM response matches with the standard human answer that it was dark, implying it was nighttime, but it falsely adds that the weather was cloudy. There was no mention or indication in the image contents that could support the claim of cloudiness. Therefore, the LMM response involves hallucination.

Rating: 1, somewhat informative, with hallucination.

LLaVA-1.5-7b (CAI):

The weather in the picture was dark, with a nighttime setting.

GPT-4 Evaluation:

The LMM's response matches the standard human-generated answer and makes no false claims about the image contents. Therefore, this response does not contain any hallucination. Both responses indicate that the weather was dark during the capturing of the photo, suggesting a nighttime setting. However, the LMM's response does not provide additional helpful insights, other than stating the darkness of the setting.

Rating: 4, somewhat informative, no hallucination.

Figure 12: Non-caption query case of LLaVA-1.5-7b on MMHal-Bench.



At what time is this photo taken?

Golden Answer:

The photo is taken during the day.

LLaVA-1.5-7b (Vanilla):

This photo is taken at night.

GPT-4 Evaluation:

The LMM's response is false because it claims that the photo is taken at night, when the standard human-generated answer states that the photo was taken during the day. The model seems to be hallucinating in this case.

Rating: 0, not informative, with hallucination.

LLaVA-1.5-7b (CAI):

This photo is taken during the day.


GPT-4 Evaluation:

"The LMM's response is accurate and matches the human-generated answer, indicating that it correctly interpreted the image. However, it doesn't provide a detailed analysis hence it isn't very informative. Therefore, there is no hallucination in the information provided by the LMM.

Rating: 5, very informative, no hallucination.

Figure 13: Non-caption query case of LLaVA-1.5-7b on MMHal-Bench.



 What color is the umbrella?

Golden Answer:
The umbrella is colorful, containing purple, yellow, teal, and orange.

LLaVA-1.5-7b (Vanilla):
The umbrella is purple.

GPT-4 Evaluation:
The LMM's response is considered as hallucination because it incorrectly identifies the umbrella as only having the color purple, while the standard human-generated answer suggests that the umbrella is colorful, containing purple, yellow, teal, and orange. This means the LMM's response has failed in providing correct, comprehensive information about the image. Therefore, the rating would be 1, somewhat informative, with hallucination.

Rating: 1, somewhat informative, with hallucination.

LLaVA-1.5-7b (CAI):
The umbrella is colorful, featuring a combination of purple, yellow, orange, and green.

GPT-4 Evaluation:
The LMM identifies the umbrella as being colorful and mentions similar colors as in the standard human-generated answer. However, it mentions green instead of teal. This is not necessarily a hallucination as teal can be seen as a shade of green. Therefore, the answer is generally in alignment with the objects listed in the image contents and it correctly answers the question.

Rating: 5, very informative, no hallucination.

Figure 14: Non-caption query case of LLaVA-1.5-7b on MMHal-Bench.