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. Previous research has shown that hallucinations are primarily caused by insufficient attention to visual information. To tackle this, recent works either depend on expensive manual annotations and computational 011 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 014 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, while preserving other foundational capabilities of LVLMs.

1 Introduction

042

With advances in computational power and data availability, large vision-language models (LVLMs) have further extended large language models (LLMs) to vision-language tasks (Liu et al., 2024a). However, 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., 2024) shows that LVLMs' insufficient attention to visual information is considered a key factor in hallucination. Recent works for mitigating hallucination either use contrastive decoding strategy (Leng



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.

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 exploring how to enhance LVLMs' perception capability by providing sufficient attention to visual information. In this work, we identify caption-sensitive attention heads and reveal a critical phenomenon: these heads significantly reduce attention to visual information when fed non-caption queries compared to caption queries. As shown in Figure 1 (a) and (b), we observe that several attention heads reduce local and global attention of the image when responding to

065

067

073

074

086

087

094

100

101

102

103

104

106

non-caption queries. In contrast, visual attention is significantly stronger when fed caption queries.

Inspired by the phenomenon of visual attention variation when fed caption compared to noncaption queries, we propose Caption-sensitive Attention Intervention (CAI), a training-free, plugand-play method, to refine caption-sensitive attention heads outputs during inference to regain 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 use the selected heads to compute attention output heads shifts, which measure how much the attention output deviates from non-caption to caption queries. Finally, we apply the precomputed attention output shifts to intervene caption-sensitive attention heads during inference, to imitate attention activation patterns when feeding LVLMs caption queries. As shown in Figure 1 (b) and (c), after using CAI, LVLM regains sufficient attention to visual information 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., 2024b), Qwen-VL-Chat (Bai et al., 2023), and LLaVA-NeXT (Liu et al., 2024c). On the POPE (Li et al., 2023) benchmark, the accuracy and the F1 score improve by 5.28% and 6.01% on average. On the MME (Fu et al., 2023) hallucination subset, the scores increase by 61.5 points on average. Furthermore, hallucination rates decrease by 6.8% on the MMHalBench (Sun et al., 2023), while the informativeness of the generated responses improves.

The main contributions can be summarized as:

(1) We propose **CAI**, a training-free, plug-andplay method significantly mitigates object hallucination in LVLMs by refining caption-sensitive attention head outputs during the inference.

(2) Compared to methods using captioning description and contrastive decoding, CAI achieves lower inference latency and better performance.

(3) Comprehensive experimental results demonstrate that CAI effectively mitigates hallucination while maintaining other foundational capabilities.

2 Related Works

2.1 Large Vision-Language Models

Recent advances in large vision-language models have significantly pushed the boundaries of multimodal understanding and generation. Several powerful LVLMs based on open-source LLM backbones combined with visual encoders, such as LLaVA-1.5-7b and Qwen-VL-Chat, have achieved impressive capabilities through extensive visionlanguage 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). Closedsource models, such as GPT-40 (Hurst et al., 2024) and Gemini 1.5 (Reid et al., 2024) have demonstrated even more powerful performance. However, current LVLMs still suffer from hallucination problems. Addressing how to efficiently and costeffectively mitigate hallucination is still a critical problem that requires further exploration.

107

109

110

111

112

113

114

115

116

117

118

119

120

121

122

123

124

125

126

127

128

130

131

132

133

134

135

136

137

138

139

140

141

142

143

144

145

146

147

148

149

150

151

152

153

154

2.2 Mitigating Hallucination in LVLMs

Current methods for mitigating hallucination in LVLMs can be broadly categorized into two methodes: training-free methods and data-driven training methods. For training-free methodes, 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). On the other hand, 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). However, there has been limited exploration of leveraging LVLMs' inherent captioning capability and multi-head attention with the goal of mitigating hallucination in LVLMs. Our work aims to address this issues by applying caption-sensitive attention intervention during the inference.

3 Methods

3.1 Task Formulation

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

155

156

157

158

160

161

162

163

164

167

168

170

171

172

174

175

176

177

178

179

180

182

183

184

186

187

188

190

191

192

193

194

195

196

During the forward pass, the input 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 $Q_{(l,h)} \in \mathbb{R}^{(m+n) \times d}$, $K_{(l,h)} \in \mathbb{R}^{(m+n) \times d}$ and $V_{(l,h)} \in \mathbb{R}^{(m+n) \times d}$, 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:

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

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

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

$$\boldsymbol{H}^{l+1} = \boldsymbol{H}^{l} + \sum_{h=1}^{H} \boldsymbol{O}_{(l,h)} \cdot \boldsymbol{W}_{o}^{l}, \qquad (2)$$

where $\boldsymbol{W}_{o}^{l} \in \mathbb{R}^{Hd \times d}$ 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 $A_{(l,h)}$ and non-caption attention weight matrix $A'_{(l,h)}$. The attention weight shift matrix can be computed as:

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

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:

$$\underset{j}{\operatorname{arg\,min}} \sum_{b=1}^{B} \boldsymbol{A}_{shift}^{b,j} \quad s.t. \ j \in J, \qquad (4)$$

199

200

201

204

205

207

208

209

211

212

213

214

215

216

217

218 219

221

222

224

227

228

229

231

232

233

234

237

where $A_{shift}^{b,j}$ 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:

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

$$\hat{\boldsymbol{O}}_{(l,h)} = \operatorname{softmax}(\dot{\boldsymbol{A}}_{(l,h)})\boldsymbol{V}_{(l,h)}, \qquad (6)$$

$$\widetilde{\boldsymbol{O}}_{(l,h)} = \widehat{\boldsymbol{O}}_{(l,h)}[m+n]. \tag{7}$$

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 noncaption query are denoted as $\widetilde{O}^{b}_{(l,h)}$ and $\widetilde{O'}^{b}_{(l,h)}$. Respectively, the last token's origin attention output are denoted as $O^{b}_{(l,h)}$ and $O'^{b}_{(l,h)}$.

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 wether the input sentence is a caption query:

$$\underset{f_{l,h(\cdot)}}{\operatorname{arg\,min}} \sum_{b=1}^{B} \mathcal{L}\left(f_{l,h}\left(x_{b}\right), y_{b}\right), \qquad (8)$$

where $x_b \in \{\widetilde{O}_{(l,h)}^b, \widetilde{O'}_{(l,h)}^b\}$ 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:

$$\boldsymbol{S}_{(l,h)} = \frac{1}{B} \sum_{b=1}^{B} \left(\boldsymbol{O}_{(l,h)}^{b} - \boldsymbol{O'}_{(l,h)}^{b} \right).$$
(9) 238



Figure 2: An overview of CAI method. Each square in the matrix represents the attention head output. Squares with dark color indicate significant output differences. 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 activating the model's inherent fine-grained visual perception to mitigate hallucination.

3.4 Intervention at Inference Time

240

241

242

244

245

246

248

249

250

251

262

This module aims to refine the top K attention heads output that are most sensitive to caption queries at inference-time. The goal is to align the attention activation pathway closer to the pre-trained caption state, thereby enhancing the model's finegrained visual perception capability and mitigate hallucination. At each layer, the updated hidden state after intervention is computed as:

$$\boldsymbol{H}^{l+1} = \boldsymbol{H}^{l} + \sum_{h=1}^{H} \left(\boldsymbol{O}_{(l,h)} + \mathbb{I}_{(l,h)} \alpha \boldsymbol{S}_{(l,h)} \right) \cdot \boldsymbol{W}_{o}^{l},$$
(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 methodes.

8 Discriminative Tasks

POPE employs a binary question-answering format, inquiring LVLMs to answer if a special object exists in the given image. This benchmark is structured around three distinct subsets: MS-COCO (Lin et al., 2014), A-OKVQA (Schwenk et al., 2022), and GQA (Hudson and Manning, 2019). 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. *objectlevel*(existence, count) and *attribute-level*(position, and color) tasks are specially designed for hallucination evaluation. 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:

$$\begin{aligned} \text{CHAIR}_{i} &= \frac{|\{\text{Hallucinationted objects}\}|}{|\{\text{All objects mentioned}\}|} & 2\\ \text{CHAIR}_{s} &= \frac{|\{\text{Sentence with hallucination objects}\}|}{|\{\text{All sentence}\}|} & 2\\ \end{aligned}$$

MMHal-Bench comprises 96 meticulously designed VQA questions, which evaluates responselevel hallucination rate and informativeness. It asks

Dataset	Setting	Method	LLaVA	LLaVA-1.5-7b		/L-Chat	LLaVA-NeXT		
2 4 4 4 5 4 4	String	1.100100	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score	
	Random	Regular VCD OPERA CAI(ours)	83.29 87.73 89.20 90.13(+6.84)	81.33 87.16 88.81 89.61(+8.28)	84.63 86.93 85.71 89.40(+4.77)	82.61 85.46 84.64 88.54(+5.93)	84.78 88.76 90.27 91.93(+7.15)	86.43 89.57 89.71 91.57(+5.14)	
MS-COCO	Popular	Regular VCD OPERA CAI(ours)	81.88 85.38 86.64 88.57(+6.69)	80.06 85.06 86.62 88.16(+8.10)	83.63 85.17 84.82 87.97(+4.34)	81.53 83.68 83.99 87.19(+5.66)	83.23 87.01 87.16 89.70(+6.47)	84.77 87.70 87.68 89.49(+4.72)	
	Adversarial	Regular VCD OPERA CAI(ours)	78.96 80.88 81.24 84.50(+5.54)	77.57 81.33 81.38 84.60(+7.03)	81.03 83.10 82.67 84.57(+3.54)	79.30 82.04 79.89 84.15(+4.85)	81.19 84.80 85.20 86.10(+4.91)	82.50 85.23 85.54 86.30(+3.80)	
	Random	Regular VCD OPERA CAI(ours)	83.45 86.15 88.02 90.40(+6.95)	82.56 86.34 84.59 90.70(+8.14)	84.93 87.33 85.77 89.70(+4.77)	83.68 86.52 84.56 89.50(+5.82)	84.20 89.27 88.20 91.93(+7.73)	84.22 88.65 87.66 92.22(+8.00)	
A-OKVQA	Popular	Regular VCD OPERA CAI(ours)	79.90 81.85 83.22 85.33(+5.43)	79.59 82.82 84.67 86.45(+6.86)	84.67 87.37 86.20 90.80(+6.13)	83.30 86.50 85.77 90.52(+7.22)	78.83 86.37 85.78 88.63(+8.00)	80.94 85.99 85.48 88.51(+7.57)	
	Adversarial	Regular VCD OPERA CAI(ours)	74.04 74.97 73.82 76.40(+2.36)	75.15 77.73 77.91 79.87(+4.72)	78.87 81.30 80.34 83.00(+4.13)	78.54 81.38 79.73 83.78(+5.24)	74.50 79.97 78.78 81.60(+7.10)	75.40 80.77 78.43 82.64(+7.24)	
GQA	Random	Regular VCD OPERA CAI(ours)	83.73 86.65 88.13 90.07(+6.34)	82.95 86.99 88.91 90.29(+7.34)	83.93 86.97 85.30 89.23(+5.30)	82.19 85.66 85.08 88.62(+6.43)	85.87 87.83 88.27 90.20(+4.33)	84.87 87.12 87.92 90.56(+5.69)	
	Popular	Regular VCD OPERA CAI(ours)	78.17 80.73 79.27 81.37(+3.20)	78.37 82.24 82.11 83.21(+4.84)	82.40 84.90 83.16 87.87(+5.47)	80.91 83.87 82.76 87.36(+6.45)	79.20 83.17 82.07 83.73(+4.53)	80.66 82.77 81.90 83.69(+3.03)	
	Adversarial	Regular VCD OPERA CAI(ours)	75.08 76.09 75.00 76.93(+1.85)	76.06 78.78 78.71 80.01(+3.95)	78.57 81.30 81.65 84.03(+5.46)	77.32 80.57 80.98 84.01(+6.69)	77.73 78.70 77.20 81.03(+3.30)	77.87 79.10 77.57 81.51(+3.64)	

Table 1: Main results on POPE tasks. We evaluate the accuracy and F1 Score of various LVLMs on the POPE tasks across the MS-COCO, A-OKVQA, and GQA datasets. **CAI(ours)** demonstrates the best hallucination mitigation performance among several methods. The comparison with more advanced methods can be found in Appendix B.

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., 2024c) as our baseline LVLMs. We compared our method with severval SOTA baselines of both *training* and *training-free* methods, including VCD, OPERA, DoLA (Chuang et al., 2023), Woodpecker (Yin et al., 2024), LURE (Zhou et al., 2023), HALC (Chen et al., 2024b), LLaVA-RLHF (Sun et al., 2023) and CCA-LLaVA (Xing et al., 2024).

Implementation Details. In our experience, 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 captionsensitive attention heads and computed the attention shift. For each attention head, SVM (Cortes, 1995) was used as the classifier and 2-fold crossvalidation was performed to evaluate its accuracy. More detailed experimental procedures are provided in Appendix A. 304

305

306

307

308

309

310

311

312

313

4.2 Main Results

Comprehensive evaluations demonstrate that our314method exhibits superior hallucination mitigation315capabilities in discriminative and generative tasks.316

303

290

Method		aVA-1.5-71	Qwen-VL-Chat							
	Existence	Count	Position	Color	Total	Existence	Count	Position	Color	Total
Greedy	170.00	121.33	115.00	152.33	558.66	175.00	140.00	123.33	170.00	608.33
VCD	180.33	131.67	125.00	155.00	592.00	180.00	133.33	131.67	175.00	620.00
OPERA	165.00	116.00	133.33	149.00	563.33	180.00	140.00	138.33	175.00	633.33
LURE	167.67	118.00	108.00	138.67	532.34	175.00	138.67	126.67	170.00	610.34
HALC	185.00	138.00	126.67	158.33	608.00	180.00	143.33	125.00	175.00	623.33
CCA-LLaVA	190.00	148.33	128.33	170.00	636.66	-	-	-	-	-
CAI(ours)	190.00	141.66	140.00	170.00	641.66	185.00	150.00	133.33	180.00	648.33

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



Figure 3: Main results on the MME full set. CAI leads to the best enhancement in LVLMs' perception capacities while preserving their recognition competencies.

Result on POPE. Table 1 presents the POPE tasks 317 results under nine different experimental settings. (1) SOTA performence. Compared with VCD and 319 OPERA, our methodes achieved superior hallucination mitigation effects across all experimental 321 configurations. Specifically, it leads to an average improvement of 5.02% in accuracy and 6.58% 323 in F1 score for LLaVA-1.5-7b, 4.87% and 6.03% for Qwen-VL-Chat, 5.95% and 5.43% for LLaVA-NeXT, resulting in SOTA hallucination mitiga-326 327 tion effects. (2) Easy to deploy in open-source LVLMs. During the best caption query searching, 328 caption-sensitive attention heads probe and shift vector computation stages, we selected images, caption query candidates and non-caption queries from 331 LLaVA-1.5-7b pre-training dataset, which are outside the domain of the benchmark datasets. De-333 spite this cross-domain discrepancy, our method 334 achieved significant improvements even when applied to Qwen-VL-Chat and LLaVA-NeXT, indi-336 cating that CAI does not rely on specific models or 337 data and deploy in open-source LVLMs easily.

Results on MME. Table 2 and Figure 3 respectively present the experimental results for the MME
hallucination subset and full set. Our method ef-

fectively 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 83.0 for LLaVA-1.5-7b and 40.0 for Qwen-VL-Chat. 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.



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.

350

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 reductions of 3.6% in the CHAIR_s metric and 1.27% in the CHAIR_i metric.

Method	LLa	VA-1.5-7b	Qwer	n-VL-Chat					
	Score↑	VH Rate%↓	Score↑	VH Rate%↓					
Training-Free Methods									
Greedy VCD OPERA	1.86 2.12 2.15	63.5 54.2 54.2	2.93 2.77 2.94	41.1 39.2 38.4					
Training Methods									
LLaVA-RLHF CCA-LLaVA	2.02 2.22	62.5 59.5	-	-					
CAI(ours)	2.33	53.0	3.04	38.0					

Table 3: Main result on MMHal-Bench.

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

5 Analysis and Discussions

5.1 Inference Latency

Method	TTFT(ms)	TPOT(ms)	Acc(%)
LLaVA-1.5-7b	99.8 1.0×	36.0 1.0×	78.96
+VCD +OPERA +VDGD	160.1 1.6× 109.8 1.1× 377.8 3.8×	96.8 2.7× 69.5 1.9× 340.9 9.5×	80.88 81.24 65.82
+CAI(ours)	102.2 1.0×	36.5 1.0×	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, achieveing faster inference speed and better performance.

5.2 Necessity of the Search Algorithm

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



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

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



Figure 6: The accuracies of classifiers.

401

402

377

352 353 354

363

364

365

370

372



Figure 7: 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.

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. These layers are critical for balancing visual perception and semantic understanding within the model. By refining the output of these attention heads, CAI significantly enhances LVLMs' visual perception capability while maintaining strong language comprehension, mitigating hallucination in the process.

5.4 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 7, CAI effectively mitigates hallucination through sentence continuation and regeneration. See the Appendix E for more cases.

5.5 Implications of Hyperparameters



Figure 8: Ablation study of α and K.

CAI method primarily relies on two key hyperparameters: the intensity of intervention α and the number of refined attention heads K. To explore their effects on performance, we performed a series of ablation experiments on the MS-COCO Adversarial POPE dataset. The results are shown in Figure 8. The key implications can be summarized as follows: 427

428

429

430

431

432

433

434

435

436

437

438

439

440

441

442

443

444

445

446

447

448

449

450

451

452

453

454

455

456

457

458

459

460

461

462

(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.

6 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. Moreover, CAI achieves lower inference latency and better performance compared to other contrastive-decoding based training-free methods. Consistent performance improvements across diverse benchmarks highlight its robustness.

403

404

405

406

407

408

409

410

411

412

413

414

7 Limitations

463

483

484

485

486

487

489

490

491

492

493

494

495

496

497

498

499

500

503

506

507

510

511

512

While CAI demonstrates significant effectiveness 464 in mitigating object hallucinations in Large Vision-465 Language Models (LVLMs), several limitations 466 should be acknowledged to provide a balanced per-467 spective on its applicability and scope. First, CAI relies on the availability of suitable caption queries 469 to identify the optimal attention shift. While our 470 search algorithm is designed to minimize this de-471 pendency, the performance of CAI may vary de-472 pending on the quality and relevance of the cap-473 tion queries used. In scenarios where high-quality 474 caption queries are unavailable, the effectiveness 475 of CAI could be limited. Expanding the caption 476 query candidate list could address this issue. More-477 over, LVLMs that do not rely on multi-head atten-478 tion mechanisms, or those employing non-standard 479 visual-textual alignment strategies, may not benefit 480 from CAI's intervention method. We will address 481 the above issues in future work. 482

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 Hussain, 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. 2024. Unveiling visual perception in language models: An attention head analysis approach. *ArXiv*, abs/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*. 513

514

515

516

517

518

519

520

521

522

523

524

525

526

527

528

529

530

531

532

533

534

535

536

537

538

539

540

541

542

543 544

545

546

547

548

549

550

551

552

553

554

555

556

557

558

559

560

561

562

563

564

565

566

- 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, Weihan 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.
- Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning

569 571

568

- 577 579 580
- 582 583

581

584

- 586

- 594

596 597

- 599

603

611 612

613

614 615 616

617 618

619

623

ings of the IEEE/CVF conference on computer vision Lee. 2024b. Improved baselines with visual instrucand pattern recognition, pages 6700-6709. tion tuning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Aaron Hurst, Adam Lerer, Adam P Goucher, Adam pages 26296–26306. Perelman, Aditya Ramesh, Aidan Clark, AJ Os-

and compositional question answering. In Proceed-

trow, Akila Welihinda, Alan Hayes, Alec Radford,

et al. 2024. Gpt-40 system card. arXiv preprint

Zhehan Kan, Ce Zhang, Zihan Liao, Yapeng Tian, Wen-

ming Yang, Junyuan Xiao, Xu Li, Dongmei Jiang,

Yaowei Wang, and Qingmin Liao. 2024. Catch: Com-

plementary adaptive token-level contrastive decoding

to mitigate hallucinations in lvlms. arXiv preprint

Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Min-

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 decod-

ing. In Proceedings of the IEEE/CVF Conference

on Computer Vision and Pattern Recognition, pages

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

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

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang,

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,

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser

Yacoob, and Lijuan Wang. 2023. Mitigating halluci-

nation in large multi-modal models via robust instruc-

tion tuning. In The Twelfth International Conference

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen,

Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024a. A survey on hallucination

in large vision-language models. arXiv preprint

models. arXiv preprint arXiv:2305.10355.

Part V13, pages 740–755. Springer.

on Learning Representations.

arXiv:2402.00253.

Wayne Xin Zhao, and Ji-Rong Wen. 2023. Eval-

uating object hallucination in large vision-language

arXiv preprint arXiv:2311.07362.

joon Seo. 2023. Volcano: mitigating multimodal

hallucination through self-feedback guided revision.

arXiv:2410.21276.

arXiv:2411.12713.

13872-13882.

arXiv:2501.01926.

Processing Systems, 36.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae

624

625

626

627

628

629

630

631

632

633

634

635

636

637

638

639

640

641

642

643

644

645

646

647

648

649

650

651

652

653

654

655

656

657

658

659

660

661

662

663

664

665

666

667

668

669

670

671

672

673

674

675

676

- 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.
- Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In European conference on computer vision, pages 146-162. Springer.
- 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, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2024a. Rlhf-v: Towards
- 10

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

678

679

681

683

684

690

691

694

695

697

700

701

702

703

704

- Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. 2024b. Rlaifv: 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*.
 - 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 visionlanguage 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*.
 - Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*.
 - Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. 2024. Ibd: Alleviating hallucinations in large vision-language models via imagebiased decoding. *arXiv preprint arXiv:2402.18476*.

780

759

760

761

762

763

764

765

766

767

768

769

770

771

772

774

775

776

778

779

781

A **Additional Experimental Details**

All datasets used in this paper are licensed under a Creative Commons Attribution 4.0 License.

A.1 Data Source

711

712

713

714

715

716

718

719

721

722

723

729

731

733

734

735

740 741

742

743

745

747

748

750

751

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 pretraining 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 752 direct sampling setting. To better compare with 754 the results of other methods, we adopted the default experimental settings of other methods. For 755 reproducibility of our results, we used greedy decoding and set $\alpha = 1.5$ and K = 100 in the main experiments. 758

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), HIO (Chen et al., 2024a), CAUSALMM (Zhou et al., 2024) and CCA-LLaVA (Xing 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.

Additional Case Studies E

More case studies are shown as follows.

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
HIO (NeurIPS'24)	90.21	89.94	88.12	87.84	84.32	84.34	87.55	87.37
CAUSALMM (ICLR'25)	88.93	88.10	87.13	87.26	83.70	82.78	86.59	86.05
CCA-LLaVA (NeurIPS'24)	88.03	86.65	86.87	85.54	85.67	84.42	86.86	85.54
CAI(ours)	90.13	89.61	88.57	88.16	84.50	84.60	87.73	87.46

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

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
VČD	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**.



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 nformativeness slightly.

Rating: 5, very informative, no hallucination.

Figure 9: 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.

O 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\u2014an 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 10: Another caption query case of LLaVA-1.5-7b on MMHal-Bench.



 \succ 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 11: Non-caption query case of LLaVA-1.5-7b on MMHal-Bench.



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.



Golden Answer: The photo is taken during the day.

LLaVA-1.5-7b (CAI):

This photo is taken during the day.

GPT-4 Evaluation:

"The LMM's response is accurate and matches the humangenerated 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 12: 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 humangenerated 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 13: Non-caption query case of LLaVA-1.5-7b on MMHal-Bench.