Cross-Modal Attention Calibration for LVLM Hallucination Mitigation

Jiaming Li[1*†] Jiacheng Zhang[2*] Zequn Jie[3] Lin Ma[3] Ming Li[4] Xiaonan Luo[5] Guanbin Li [1,6,7‡]
[1] Sun Yat-sen University [2]The University of Hong Kong [3]Meituan
[4] Inspur Database Technology [5] Guilin University of Electronic Technology
[6] Shenzhen Loop Area Institute [7]Guangdong Key Laboratory of Big Data Analysis and Processing
lijm48@mail2.sysu.edu.cn, liguanbin@mail.sysu.edu.cn

Abstract

Large vision-language models (LVLMs) have shown remarkable capabilities in visual-language understanding. Despite their success, LVLMs still suffer from generating hallucinations in complex generation tasks, leading to inconsistencies between visual inputs and generated content. To address this issue, some approaches have introduced inference-time interventions, such as contrastive decoding, to reduce overreliance on language priors. However, these approaches overlook hallucinations stemming from position bias and spurious inter-modality correlations. In this paper, we propose a Cross-Modal Attention Calibration (CMAC) method to mitigate hallucinations in LVLMs in a training-free manner. In this method, we design an Inter-Modality Decoding (IMD) module to alleviate hallucination by a novel contrastive decoding mechanism. IMD masks the value vectors associated with significant cross-modal attention weights as distortion, which addresses both unimodality overreliance and misleading inter-modality correlations. Additionally, a Cross-Modal Position Calibration (CMPC) module shrinks the position gap of image tokens, alleviating the position bias in cross-modal attention. Experimental results on diverse hallucination benchmarks validate the superiority of our method over existing state-of-the-art techniques in reducing hallucinations for LVLM. Our code will be available at `https://github.com/lijm48/IMCCD`.

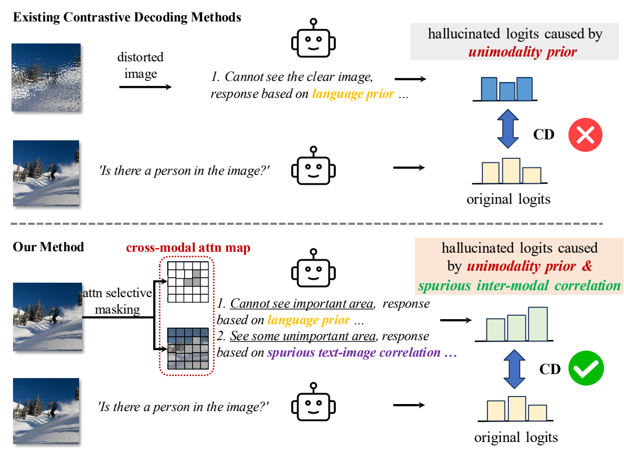


Figure 1. Illustration of the comparison between the commonly used contrastive decoding strategy method and our proposed *Inter-Modality Contrastive Decoding*.

1. Introduction

With advances in computational power and data availability, large language models [1, 3, 11, 35, 36] (LLMs) have achieved significant progress in language understanding[23, 24], generation[15], and reasoning[37]. Large vision-language models [4, 6–8, 29, 30] (LVLMs) further extend large language models to vision-language tasks, demonstrating impressive performance across a range of applications, including image captioning and visual question answering (VQA). Despite these advancements, LVLMs suffer from the issue of hallucinations while generating the response, in which LVLMs generate textual content that is semantically coherent but inconsistent with ground-truth objects in the given image, hindering their reliable application.

Among various mitigation strategies, inference-time methods, particularly Contrastive Decoding (CD)[13, 21, 40], have gained prominence due to their training-free nature and effectiveness. The core principle of CD is to penalize tokens that are likely under a ”hallucinatory” distribution, typically induced by perturbing the input (e.g., distorting visual features or masking text). While promising, we argue that existing CD-based approaches suffer from two fundamental, previously overlooked limitations. First, they fail to disentangle genuine visual grounding from spurious cross-modal correlations. LVLMs often learn superficial as-

[*]Equally-contributed authors.
[†]Work done during an internship at Meituan.
[‡]Corresponding author is Guanbin Li.

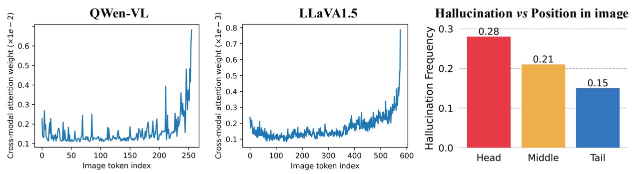


Figure 2. Illustration of the cross-modal attention *Position Bias* within some common LVLMs and the resulting spatially biased hallucination patterns.

sociations from web-scale data, such as the co-occurrence of ”food” and ”dining table” [20]. Consequently, a query about a dining table can trigger a hallucinated response even if one is absent, merely because food is present in the image. Current CD techniques, which distort the entire visual input, cannot isolate and penalize these spurious associations. The resulting ”hallucinatory” distribution is too coarse, failing to provide the fine-grained contrast needed to suppress such errors. Second, we identify a systematic positional bias in the cross-modal attention mechanism of contemporary LVLMs. As illustrated in Figure 2, we observe that hallucinations are more likely to involve objects corresponding to the initial portion of the image token sequence. With further investigation, we reveal that the issue arises from the Rotary Position Embedding (RoPE) in the LVLM decoder. Typically, images are embedded into tokens, interleaved with text tokens, and processed with RoPE for autoregressive generation. This architecture inherently incentivizes the model to disproportionately attend to the latter parts of the visual sequence during autoregressive generation, which is demonstrated by the relatively higher attention weights of the later part of image tokens.

To overcome these challenges, we introduce the **C**ross-**M**odal **A**ttention **C**alibration (CMAC) framework for more effective hallucination mitigation in a training-free manner. To address spurious correlations, an *Inter-Modality Decoding (IMD)* module is proposed to craft a more effective ”hallucinatory” distribution for contrastive decoding. As illustrated in Figure 1, instead of globally distorting the image, *IMD operates at the **attention level***. It selectively masks the value vectors corresponding to high cross-modal attention scores. This surgical intervention suppresses strong, legitimate inter-modality correlations in cross-modal attention while leaving weak, spurious correlations and the unimodal reliance intact in the contrastive distribution. This allows the resulting contrastive distribution to effectively capture failures from both spurious correlations and unimodal language priors. To counteract positional bias, we propose the *Cross-Modal Position Calibration (CMPC)* module, which mitigates the tendency to overlook crucial visual tokens caused by position bias. This is achieved by linearly reducing the positional index gap of image tokens in cross-modal attention, preserving their global position within sentences and alleviating positional bias. Extensive experiments on various LVLM hallucination benchmarks demonstrate the superiority of our CMAC framework, significantly outperforming existing methods. Our main contributions are summarized as follows:

- We propose a novel Inter-Modality Decoding method for LVLM hallucination mitigation, which eliminates both uni-modality overreliance and spurious inter-modality correlations by contrastive decoding.
- We design a Cross-modal Position Calibration, which alleviates the position bias by scaling the position indexes of image tokens in the cross-modal attention process.
- Comprehensive experiments on various benchmarks demonstrate the effectiveness and generalization of our proposed method in reducing hallucinations for LVLMs.

2. Related Work

Large Vision-Language models. Large vision-language models (LVLMs) equip large language models (LLMs) with the capability to perceive and understand both textual input and visual input data. The most common practice to achieve LVLMs is to integrate pre-trained LLMs with additional visual encoders and cross-modal interfaces for cross-modal fusion. For example, the LVLMs [4, 6–8] represented by InternVL [8] introduce linear projections to map the image features from the vision encoder to the token space of LLMs. While LLaVA series [29, 30] develop a vision-language model by connecting a vision encoder and an LLM with a projection layer and fine-tuning their model based on their generated instructional vision-language data. Instead, another line of work [22, 43, 44, 53] such as BLIP-2 [22] adopts the design of query transformer as the interface between vision encoders and LLMs. Dai et al. [12] further proposes InstructBLIP, which enhances visual comprehension through vision-language instruction tuning and introduces instruction-aware visual feature extraction to query transformers to enable context-relevant processing of visual content based on the given instructions. Despite these advancements in LVLMs, these models continue to struggle with severe hallucination issues, where generated content misaligns with the visual inputs. Our work aims to mitigate the hallucination of current LVLMs and facilitates the application of LVLMs in various domains.

Hallucination in LVLMs. The hallucination problem was first discovered in the field of LLMs. It refers to the misalignment between the generated content from LLMs and real-world facts (namely the factuality hallucination) or user instructions (namely the faithfulness hallucination). Built upon LLMs, LVLMs also suffer from the hallucination that manifests as a misalignment between generated text and the visual input. Various approaches have been proposed to tackle this issue, including the works from the perspective of constructing additional robust training instruction [18, 28, 46, 49], reinforcement learning with human/AI feedback [14, 19, 25, 33, 42, 47, 48], or model structure enhancement [50], etc. Despite achieving impressive results, these methods always require extensive data collection or additional fine-tuning of LVLMs, making them both computationally and labor-intensive. Another line of method [2, 5, 16, 31, 39, 41, 45, 52] focuses on the training-free method by improving the inference process of LVLMs to suppress hallucination. For example, some methods [9, 13, 17, 21, 34, 38, 40] mitigate the hallucination via reducing the over-reliance of LVLMs on the language prior by performing conservative decoding on the original inputs and the inputs with disturbed contents. To the end, other methods like VCD [21] propose to distort the visual content by noise-adding [21] or token-wise pruning [13, 40], while ICD [38] implicitly distorts the instruction by introducing the negative role prefix for LVLMs. However, the distortion of textual/visual content intervenes in cross-modal attention, neglecting hallucinations caused by spurious inter-modality correlations. Additionally, the position bias is also under-explored in the training-free manner.

3. Methodology

3.1. Preliminaries

Modern LVLMs typically comprise a visual encoder, a cross-modal interface, and a language decoder. The visual encoder is usually adapted from a pre-trained vision model, while the language decoder is derived from a pre-trained LLM. During inference, LVLMs process both visual and textual inputs to generate the next token in the response sequence iteratively. Specifically, the visual encoder first encodes the input visual content into visual features. The cross-modal interface then maps these visual features to the input space of the language decoder to generate the image tokens $\mathbf{X} = \{x_i\}_{i=0}^{n}$. Here x_i is the image token corresponding to the i-th patch of the image and n is the number of patches. Besides, the text input is mapped to the text tokens $\mathbf{T} = \{t_i\}_{i=0}^{m}$, where t_i is the i-th text token and m is the number of text tokens. The image tokens and text tokens are then concatenated to serve as the input tokens $[\mathbf{T}_{0:m_b}, \mathbf{X}, \mathbf{T}_{m_b+1:m}]$ for the language decoder. $\mathbf{T}_{0:m_b} = \{t_i\}_{i=0}^{m_b}$ is the first m_b tokens (system prompt) in $\mathbf{T}$ and $\mathbf{T}_{m_b+1:m} = \{t_i\}_{i=m_b+1}^{m}$ is the remaining part of $\mathbf{T}$ (query prompt).

Language Decoder Forwarding. The multi-head self-attention layer is commonly used in the language decoder. During the forward pass of the language decoder, the input tokens are initially transformed into input embeddings by the embedding layer, which then serve as the hidden states for the first self-attention layer. For each sample, every head within a self-attention layer maps the hidden states to queries $\mathbf{Q} \in \mathbb{R}^{(n+m)\times d}$, keys $\mathbf{K} \in \mathbb{R}^{(n+m)\times d}$, and values $\mathbf{V} \in \mathbb{R}^{(n+m)\times d}$ by linear transformation. Here $(n+m)$ is the sequent length of input tokens and d denotes the hidden dimensions. Then the position embeddings are applied to the queries and keys to incorporate the position information of each token. Recent language decoders primarily rely on the rotary position embeddings (ROPE) to encode the position information, which estimate rotary matrices $\mathbf{R}$ for query and key states by,

$$\mathbf{R}_k^i = \begin{pmatrix} \cos(\mathbf{P}[i]\theta_k) & -\sin(\mathbf{P}[i]\theta_k) \\ \sin(\mathbf{P}[i]\theta_k) & \cos(\mathbf{P}[i]\theta_k) \end{pmatrix}, \quad \theta_k = \frac{1}{b^{2k/d}}$$
$$\mathbf{q}_{\mathrm{p}}^i = \mathbf{q}^i\mathbf{R}^i, \quad \mathbf{k}_{\mathrm{p}}^i = \mathbf{k}^i\mathbf{R}^i, \tag{1}$$

Here b is a hyper-parameter. $\mathbf{R}_k^i$ is the rotary matrix for the $2k$-th and $(2k+1)$-th dimension of query $\mathbf{q}^i$ and key states $\mathbf{k}^i$ of i-th token with position index $\mathbf{P}[i]$. Normally, the position indexes of all the tokens are set to $\mathbf{P} = [\{i\}_{i=1}^{m+n}]$. Given the queries $\mathbf{Q}_{\mathrm{p}}$ and keys $\mathbf{K}_{\mathrm{p}}$ with position embeddings, the attention matrix $\mathbf{A} \in \mathbb{R}^{(n+m)\times(n+m)}$ is estimated by,

$$\mathbf{A}^l = \frac{\mathbf{Q}_{\mathrm{p}}\mathbf{K}_{\mathrm{p}}{}^T}{\sqrt{d}}, \quad \mathbf{A} = \mathrm{softmax}(\mathbf{A}^l), \tag{2}$$

where $\mathbf{A}^l$ is the attention logits before softmax. The attention matrix $\mathbf{A}$ estimates the relevance of each token, which is used to reweight the values $\mathbf{V}$ from each token to obtain the attention output $\mathbf{O} \in \mathbb{R}^{(n+m)\times d}$,

$$\mathbf{O} = \mathbf{A}\mathbf{V}, \tag{3}$$

The attention output $\mathbf{O}$ is subsequently fed into the fully connected feed-forward network (FFN), whose output then serves as the input hidden states for the subsequent multi-head self-attention layer.

Next Token Prediction. The language decoder parameterized by θ in LVLMs generates the t-th output token based on the text tokens $\mathbf{T}$, image tokens $\mathbf{X}$, and the previously generated tokens $\mathbf{Y}_{<t}$ in an auto-regressive manner. This process can be formulated as follows,

$$\begin{aligned} \mathbf{Y}_t &\sim p_\theta(\mathbf{Y}_t|\mathbf{T}, \mathrm{Attn}(\mathbf{O}|\mathbf{T}, \mathbf{X}, \mathbf{Y}_{<t})), \\ &\propto \mathrm{logit}_\theta(\mathbf{Y}_t|\mathrm{Attn}(\mathbf{O}|\mathbf{T}, \mathbf{X}, \mathbf{Y}_{<t})), \end{aligned} \tag{4}$$

where $p_\theta(\cdot)$ and $\mathrm{logit}_\theta(\cdot)$ denote the output probability scores and logits of the language decoder, respectively. Attn is the self-attention process by Eq (2) and Eq (3). The output token $\mathbf{Y}_t$ is subsequently concatenated with previously generated tokens to serve as input for the next step in the token generation sequence.

Contrastive Decoding for LVLMs. Contrastive Decoding [26] is an effective training-free intervention strategy for suppressing hallucinations in LVLMs. Its core idea is to introduce a manually crafted, distorted input to induce the model to generate hallucinated outputs, which are then used to mitigate potential hallucinations in the original output by contrasting original output distribution logits to the distorted counterpart. Existing works propose designing distorted inputs—either textual [51] or visual [21]—to counteract *hallucinations caused by uni-modal priors*, where the model generates results based solely on one modality while ignoring the other. Formally, contrastive decoding can be expressed as:

$$\mathbf{l}_t = \text{logit}_\theta(\mathbf{Y}_t|\text{Attn}(\mathbf{O}|\mathbf{T}, \mathbf{X}, \mathbf{Y}_{<t})), \tag{5}$$

$$\widetilde{\mathbf{l}}_t = \text{logit}_\theta(\mathbf{Y}_t|\text{Attn}(\mathbf{O}|\widetilde{\mathbf{T}}, \widetilde{\mathbf{X}}, \mathbf{Y}_{<t})), \tag{6}$$

$$p_\theta(\mathbf{Y}_t|\mathbf{T}, \mathbf{X}) = \text{softmax}((1+\alpha)\mathbf{l}_t - \alpha\widetilde{\mathbf{l}}_t). \tag{7}$$

α is the hyper-parameter controlling the contrastive effect, $\widetilde{\mathbf{T}}$ and $\widetilde{\mathbf{X}}$ are the distorted text and visual input, respectively. Despite widely studied [9, 17, 21, 38, 40], existing methods primarily focus on the unimodal distortion, neglect the spurious inter-modality correlations, leading to suboptimal estimation of distorted distribution.

3.2. IMCCD for LVLMs Hallucination Mitigation

In this section, we present the Cross-Modal Attention Calibration (CMAC) framework, designed to mitigate hallucinations in LVLMs. CMAC incorporates two key components: Inter-Modality Decoding (IMD) and Cross-Modal Position Calibration (CMPC), which are used to address the hallucination caused by the spurious inter-modality correlations and positional biases, respectively. An overview of our framework is illustrated in Figure 3.

3.2.1. Inter-Modality Contrastive Decoding

Existing contrastive decoding methods primarily address the hallucinations caused by over-reliance on the uni-modal priors(e.g., over-reliance on either text or vision) through input distortions, while neglecting *hallucinations arising from spurious inter-modality correlations*. To bridge this gap, we propose Inter-Modality Decoding (IMD), which corrects cross-modal correlations through targeted distortion of cross-modal attention patterns. Our key insight is to selectively suppress the value vectors corresponding to high values in attention logits and exacerbate the hallucination from the unreliable inter-modality interaction, thus subsequently alleviating it with contrastive decoding.

Specifically, we firstly isolate the cross-modal segment from the attention logs $\mathbf{A}^l \in \mathbb{R}^{(n+m)\times(n+m)}$, which is denoted as $\mathbf{A}^l_{\text{cross}} = \mathbf{A}^l[m_b+n : m+n, m_b : m_b+n]$. Generally, $\mathbf{A}^l_{\text{cross}}$ represents the attention weights between the subsequent text tokens and image tokens, capturing the inter-modality correlations within the self-attention layer. Based on such cross-modal attention weight, we aim to estimate the distribution that favors the hallucination, which afterward is used to mitigate the hallucination from spurious inter-modality correlations. To accomplish this, IMD first generates a mask that selectively identifies the prominent attention weights based on their magnitude,

$$\mathbf{M}_{\text{cross}} = \mathbb{I}(\mathbf{A}^l_{\text{cross}} > \mu(\mathbf{A}^l_{\text{cross}})), \tag{8}$$

where $\mu(\cdot)$ represents the mean operation and $\mathbb{I}$ is the indicator function. The binary mask $\mathbf{M}_{\text{cross}} \in \mathbb{R}^{(m-m_b)\times(n)}$ adaptively indicates the selected significant attention weights without the position bias for the cross-modal segment. We subsequently pad the $\mathbf{M}_{\text{cross}}$ to a global mask $\mathbf{M} \in \mathbb{R}^{(n+m)\times(n+m)}$ which applies to all attention weights in $\mathbf{A}$ with zero padding. Specifically, $\mathbf{M}[m_b+n : m+n, m_b : m_b+n]$ is set to the $\mathbf{M}_{\text{cross}}$, ensuring that the cross-modal part is accurately masked in the global context.

With this global mask, IMD then distorts the cross-modal segment of self-attention. While recent approaches typically adapt token-wise pruning image tokens, IMD masks the value vector $\mathbf{V}$ to $\mu(V)$ by dim-wise mean operation to exacerbate the hallucination. With this process, the content of the value vector is teased apart. IMD then estimates the attention output by performing the weighted sum between the origin value vector $\mathbf{V}$ and distorted value vector $\mu(\mathbf{V})$ guided by both the mask $\mathbf{M}$ and attention weights $\mathbf{A}$. Specifically, IMD modifies the process Eq (3) by the following process,

$$\widetilde{\mathbf{O}} = (\mathbf{M}\cdot\mathbf{A})\mu(\mathbf{V}) + ((1-\mathbf{M})\cdot\mathbf{A})\mathbf{V}. \tag{9}$$

With the above equation, IMD distorts the value vectors corresponding to the significant cross-modal attention weights while keeping the value vectors of the remaining attention part unchanged, which suppresses important inter-modality correlations from cross-modal attention. Similar to Eq (7), the final contrastive decoding process is performed by the following equations,

$$\mathbf{l}_t = \text{logit}_\theta(\mathbf{Y}_t|\text{Attn}(\mathbf{O}|\mathbf{T}, \mathbf{X}, \mathbf{Y}_{<t})), \tag{10}$$

$$\widetilde{\mathbf{l}}_t = \text{logit}_\theta(\mathbf{Y}_t|\text{Attn}(\widetilde{\mathbf{O}}|\mathbf{T}, \mathbf{X}, \mathbf{Y}_{<t})), \tag{11}$$

$$p_\theta(\mathbf{Y}_t|\mathbf{T}, \mathbf{X}) = \text{softmax}((1+\alpha)\mathbf{l}_t - \alpha\widetilde{\mathbf{l}}_t). \tag{12}$$

Compared with recent contrastive decoding methods, the advantages of IMD are threefold: (1) The distortion process used in IMD directly focuses on the cross-modal segment of attention to alleviate the cross-modal inconsistency. Specifically, it proposes to modify the value vectors with high attention weights on cross-modal attention, leaving the low-weight inter-modality correlations and intra-modality knowledge exchange unaffected. This approach enables IMD to address hallucinations from both uni-modality over-reliance and spurious inter-modality correlations more effectively. (2) Traditional contrastive decoding methods alter attention weights from visual tokens to text tokens, which

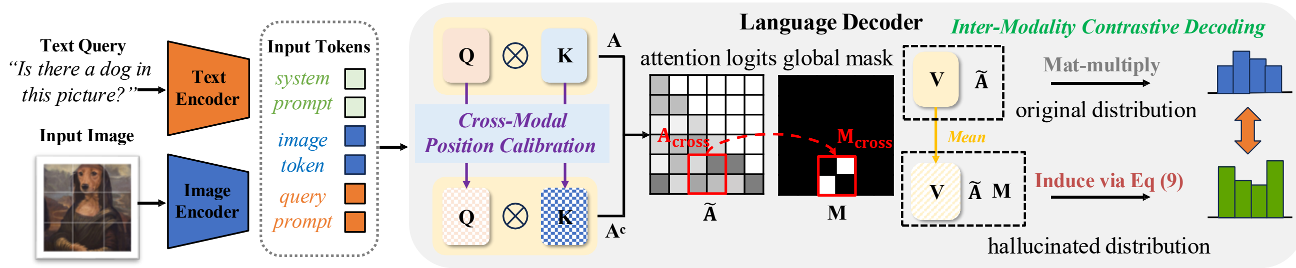


Figure 3. An overview of the proposed CMAC approach, consisting of two modules: Inter-Modality Decoding (IMD) and Cross-Modal Position Calibration (CMPC). During inference, IMC generates a distorted output distribution that favors hallucination by selectively masking value vectors corresponding to high attention weights in the cross-modal segment of the attention matrix. IMC then performs contrastive decoding between the original outputs and distorted outputs to mitigate the hallucination. Additionally, CMPC scales the position indexes to estimate the content-driven attention weights to replace the cross-modal segment, alleviating the position bias.

can result in over- or under-estimated hallucinations from image regions. IMD, however, avoids modifying attention weights, yielding a more precise estimation of distributions prone to hallucination. (3) Besides, since the attention process of the visual part remains the same, the key and value vectors and attention weights can be directly derived from the original forward process, leading to a faster inference time than other contrastive decoding methods.

3.2.2. Cross-Modal Position Calibration

Although IMD reduces hallucinations arising from the cross-modal interaction, MLLMs still face challenges with visual content retention. Specifically, with the position embeddings, the text tokens will be biased to pay more attention to the latter part of the image visual tokens, while overlooking the head part of the visual tokens, leading to more hallucinations. Motivated by this, we propose a cross-modal position calibration module to encourage the self-attention layer to alleviate the position bias from position embeddings on the cross-modal part. Concretely, we estimate new position indexes by scaling the position indices of all image tokens. It is achieved by replacing the original position indexes of concatenated input tokens, noted with $\mathbf{P} = [\{i\}_{i=1}^{m+n}]$, with refined position indexes,

$$\mathbf{P}^c = [\{i\}_{i=1}^{m_b}, \{m_b + \frac{i}{\gamma}\}_{i=1}^{n}, \{i + \frac{n}{\gamma}\}_{i=m_b+1}^{m}], \tag{13}$$

where γ is the scaling factor.

The refined position indexes shrink the gap between rotary angles of image tokens in rotary matrices in Eq (1). This keeps the global positional relation of the image content in the input language and reduces the positional bias of the head part of image tokens with other language tokens. The modified position indexes $\mathbf{P}^c$ are used to perform ROPE to estimate the attention logits $\mathbf{A}^c$ with refined position embeddings by Eq (1) and Eq (2). Then we refine the cross-modal part of attention logits $\mathbf{A}^l$ with,

$$\widetilde{\mathbf{A}}_{i,j}^l = \begin{cases} \mathbf{A}_{i,j}^c, & \text{if } j > m^b + n \text{ and } m^b + n > i \geq m^b, \\ \mathbf{A}_{i,j}^l, & \text{otherwise.} \end{cases}$$

The $\widetilde{\mathbf{A}}^l$ is used to perform the further estimation of values in Eq (8), Eq (3) and Eq (9). Note that this refinement is applied to both the estimation of logits from original inputs and distorted inputs. With CMPC, the language decoder is encouraged to prioritize image content over token positions, effectively alleviating hallucinations caused by the overlook of visual content. We provide some further analysis about the position bias of RoPE in the supplementary document.

4. Experiments

4.1. Datasets and Evaluation Metrics

POPE. POPE [27] designs a new metric and benchmark to assess object hallucination in the VQA paradigm. It achieves the evaluation of object hallucination in a binary classification task by prompting LVLMs to answer yes or no for short questions about the existence of probing objects (e.g., Is there a car in the image?). The POPE benchmark consists of nine different subsets from three datasets, MS-COCO, AOKVQA, and GQA, with three different negative sampling strategies (Random, Popular, and Adversarial). Each sampled subset includes 500 images with 6 problems for each image. Following previous works [20, 38], we add a " Please answer in one word." constraint for evaluating the POPE dataset at the end of the query text. F1-score and Accuracy are commonly reported as the metrics for the performance of hallucination mitigation on this benchmark.
Chair. The CHAIR [32] is a widely used metric for assessing the hallucination in responses of LVLMs. The CHAIR metric comprises two variants, namely the CHAIR_i and CHAIR_s, which are formulated as follows:

$$\text{CHAIR}_i = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all objects mentioned}\}|},$$

$$\text{CHAIR}_s = \frac{|\{\text{sentences with hallucinated object}\}|}{|\{\text{all sentences}\}|}.$$

Besides, we also report the recall and F1-score to show the completeness of the generated caption of the image. Following previous work[16], we randomly sample 500 images in the validation set of MS-COCO and conduct the evaluation of CHAIR metrics on the sampled subset with the prompt "Please describe this image in detail".

MME. The MME dataset is a comprehensive benchmark for the evaluation of LVLMs. Similar to POPE, these tasks are formulated as binary classifications with yes-no questions. It encompasses 14 subtasks for the examination of the perception and cognition abilities of LVLMs. The score is estimated by the sum of the accuracy of each question and the accuracy of each image. For hallucination evaluation, we report the results on both the MME full set and the hallucination subset with four hallucination-related sub-tasks, including object existence, count, position, and color.

4.2. Models and Implementation Details

We follow VCD[21] to integrate our proposed method with three popular LVLMs, LLaVA1.5 [29], InstructBLIP [12] and QWEN-VL [4]. All these models adopt Vicuna 7B [10] as their language models. We empirically set $\gamma = 2$ and $\alpha = 3$. On all the sampling setting, top p is set to 1. We leave other hyper-parameters the same with VCD.

4.3. Experimental Results

Comparison on POPE. We compare our method against three existing contrastive decoding methods, including VCD [21], ICD [38], OPERA [16] and PAI [31], on the POPE dataset. In the POPE dataset, we report the average of performance on COCO, AOKVQA, and GQA. We also follow VCD to report the performance of different subsets of POPE distinctly, which is shown in our supplementary document. The results are presented in Table 1. It can be observed that our method boosts the performance of LVLMs in mitigating hallucination from object existence for each setup of different subsets on POPE. Our proposed CMAC method outperforms the existing methods by a clear margin on different LVLMs. Compared to the second-best method PAI, CMAC achieves a notable increase in accuracy and F1 score, underscoring its effectiveness in reducing spurious inter-modality correlations. For greedy search, the LVLMs are more sensitive to distribution so some methods may leads to performance degradation. Nevertheless, the results showcases our CMAC also can achieve significantly better results with greedy search, which demonstrates the generalization capability of CMAC.

Comparison on CHAIR. We conduct evaluation of the CHAIR metric on the MS-COCO validation set in Table 2 to validate the performance of our method on the image captioning task and long sequence generation. The table shows that the performance of our method is better than that of other existing contrastive decoding methods. Specifically, our CMAC significantly enhances the performance of LVLMs on CHAIR_i and CHAIR_s by 9.6% and 5.1% for LLaVA 1.5 and by 6.4% and 3.5% for InstructBLIP under nucleus sampling, demonstrating the capability of our method for generating long sequence response with less hallucination. Besides, our CMAC also achieves a better F1 score, which showcases that the generated responses from our method can describe the image content more accurately and completely.

Comparison on MME. To validate the effectiveness of our method in mitigating various types of hallucination beyond object existence, we perform the comparison on the MME hallucination subset and the MME full set with nucleus sampling. The results are averaged over 4 runs. For MME hallucination set shown in Table 3, CMAC achieves better scores on mitigating both object-level and attribute-level hallucination. Beyond the object existence, CMAC leads to an improvement of 17.50, 8.50, and 5.42 on count, position, and color for LLaVA1.5, which demonstrates the generalization capability of CMAC across various types of hallucinations.

For the MME full set, Figure. 4 illustrates the performance of LLaVA 1.5 with normal decoding, VCD, and our method. Our method generally outperforms the other methods in 10 sub-tasks of 14. Notably, our method demonstrates a strong capability for perception, outperforming VCD and ICD by a large margin. However, our approach is less effective in numerical calculations and text translation tasks, as these primarily rely on the language decoder's reasoning capability rather than visual content comprehension. Nevertheless, our method maintains comparable performance in recognition tasks compared to existing approaches. Overall, these findings illustrate that our method enhances the general functionality of LVLMs while also effectively mitigating hallucinations.

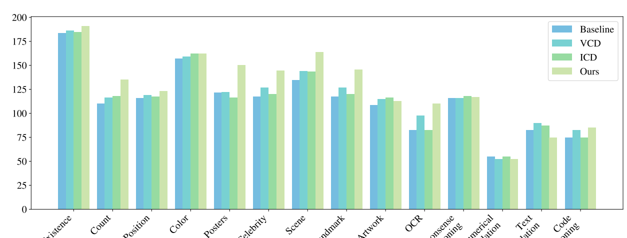


Figure 4. Results of LLaVA1.5 on MME-Fullset.

4.4. Ablation Study and Analysis

In this section, we conduct ablations to demonstrate the effectiveness of our method. Without specification, our ablation study is done on the popular setups of POPE and the MS-COCO validation set for CHAIR.

		Sampling						Greedy					
		Random		**Popular**		**Adversarial**		**Random**		**Popular**		**Adversarial**	
	Method	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
LLaVA-1.5	Baseline	83.49	82.28	79.98	79.34	76.03	76.26	83.51	82.27	79.68	79.07	75.91	76.11
	ICD	85.70	84.43	81.61	81.46	78.66	78.57	87.82	86.73	84.72	83.94	80.98	80.78
	VCD	86.84	86.83	82.65	83.37	77.31	79.28	88.11	88.15	84.23	84.91	78.08	80.07
	PAI	87.73	87.65	83.45	84.21	78.36	78.53	88.32	88.01	84.69	85.61	79.18	80.38
	OPERA	-	-	-	-	-	-	88.85	88.67	82.77	83.40	79.16	80.93
	Ours	**89.10**	**88.60**	**86.0**	**85.70**	**81.41**	**81.84**	**89.33**	**88.89**	**86.33**	**86.00**	**82.02**	**82.42**
InstructBLIP	Baseline	80.41	80.94	76.09	77.65	72.37	75.42	87.44	86.80	81.85	82.58	77.66	79.21
	ICD	85.78	85.73	81.12	82.25	76.82	78.99	86.76	86.83	80.54	82.16	76.73	78.83
	VCD	84.11	84.13	79.96	80.80	76.32	78.08	86.66	86.37	81.73	82.29	78.00	79.45
	PAI	84.60	84.54	80.13	80.94	77.00	78.33	86.51	86.19	81.65	82.34	77.76	79.13
	OPERA	-	-	-	-	-	-	84.57	83.74	78.24	79.15	74.59	76.33
	Ours	**86.67**	**86.10**	**82.20**	**82.37**	**78.37**	**79.25**	**87.68**	**87.36**	**82.38**	**82.80**	**78.77**	**79.58**
Qwen-VL	Baseline	84.12	82.42	81.89	80.51	79.10	78.07	84.09	82.77	83.97	82.63	81.40	80.38
	ICD	86.45	85.80	83.41	83.27	81.04	81.33	86.49	85.86	83.80	83.71	80.96	81.27
	VCD	87.82	87.68	85.60	85.48	81.85	82.34	87.83	87.39	85.91	85.78	81.90	82.44
	PAI	88.15	87.94	85.73	85.51	81.98	82.23	87.86	87.63	85.73	86.62	82.06	82.44
	OPERA	-	-	-	-	-	-	83.05	83.20	81.40	81.89	76.73	78.31
	Ours	**88.61**	**88.26**	**86.67**	**86.34**	**82.68**	**83.05**	**88.92**	**88.48**	**86.92**	**86.55**	**83.01**	**83.47**

Table 1. **Results on discrimination hallucination benchmark POPE.** The Baseline method denotes the standard decoding. The best performances within each setting are **bolded**. "Acc" and "F1" denote the Accuracy and F1 scores, respectively. The results are copied from the papers or from our re-implementation based on official codes.

Method	**LLaVA 1.5**				**InstructBLIP**				**Qwen-VL**			
	$\mathbf{C}_i \downarrow$	$\mathbf{C}_s \downarrow$	**Recall** $\uparrow$	**F1** $\uparrow$	$\mathbf{C}_i \downarrow$	$\mathbf{C}_s \downarrow$	**Recall** $\uparrow$	**F1** $\uparrow$	$\mathbf{C}_i \downarrow$	$\mathbf{C}_s \downarrow$	**Recall** $\uparrow$	**F1** $\uparrow$
Sampling	55.6	17.8	72.4	77.0	57.0	16.7	70.4	76.3	44.8	11.3	74.6	81.1
VCD	54.2	16.4	76.7	80.0	53.6	15.6	**74.6**	79.2	47.2	12.1	**76.1**	81.6
ICD	54.0	15.4	**77.2**	80.7	**49.6**	14.5	72.6	78.5	43.2	11.2	75.0	81.5
Ours	**47.0**	**12.7**	75.6	**81.0**	50.6	**13.2**	73.7	**79.7**	**41.2**	**10.6**	75.4	**81.8**
Greedy	49.6	15.1	77.9	81.2	48.2	13.3	73.8	79.7	45.8	11.4	74.6	81.0
VCD	55.2	15.7	77.6	80.8	49.8	14.0	**74.9**	80.1	46.6	12.2	**76.3**	81.7
ICD	57.0	15.4	**77.5**	80.9	**44.6**	12.6	73.9	80.1	42.6	11.1	74.9	81.3
Ours	**47.6**	**12.5**	76.2	**81.5**	45.2	**12.1**	74.2	**80.5**	**41.4**	**10.9**	75.6	**81.8**

Table 2. **Results on hallucination metric CHAIR on the MS-COCO validation set.** The best performances within each setting are **bolded**. $\mathbf{C}_i$ and $\mathbf{C}_s$ denotes the CHAIR$_i$ and CHAIR$_s$, respectively.

Impact of individual proposed modules. We conducted an ablation study to assess the effectiveness of each individual module, including IMD and CMPC, within our CMAC framework. As illustrated in Table 4, IMD significantly improves the performance by 5.12% on Accuracy and 5.34% on F1 for VQA task, underscoring the significance of IMD on mitigating hallucination. The combination of IMD and CMPC demonstrates a 1.02% performance improvement of F1, which validates the effectiveness of CMPC in alleviating position bias. For image-captioning tasks, IMD boosts the performance on both CHAIR$_s$. Based on IMD, CMPC slightly enhances the performance, showcasing the efficacy of CMPC for long sequence generation tasks.

Ablation of distortion used in our method. In Table 5, we show the performance of our method with different kinds of distortions. Specifically, 'Attention mask' indicates pruning significant attention weights by the conventional attention mask strategy. 'Value noise addition' indicates adding noise on value vectors similar to VCD [21] instead of masking the value vectors to the mean value, which is denoted as 'Value mask'. Our method shows a better hallucination mitigation capability compared with attention masking, which demonstrates that altering the attention weights may lead to over- or under-estimation of hallucinations from image regions. In contrast, the value mask distortion in our method also surpasses the value noise addition, achieving a better

Model	Decoding	Object-level *Existence*↑	*Count*↑	Attribute-level *Position*↑	*Color*↑	Total↑
LLaVA1.5	Baseline	183.75	110.00	115.83	157.08	566.67
	VCD	186.25	116.67	119.17	159.17	581.25
	ICD	185.00	117.91	117.50	162.08	582.50
	Ours	**191.25**	**135.41**	**123.33**	**162.50**	**612.49**
InstructBLIP	Baseline	154.17	**86.67**	58.33	123.75	422.91
	VCD	166.67	82.08	67.08	130.00	445.83
	ICD	166.67	80.83	67.92	128.33	443.75
	Ours	**183.75**	85.00	**68.75**	**131.67**	**469.17**
Qwen-VL	Baseline	142.50	116.67	134.08	157.08	550.33
	VCD	148.58	113.67	133.00	164.92	560.17
	ICD	144.75	116.44	**135.17**	162.75	559.11
	Ours	**157.75**	**118.75**	132.83	**168.75**	**578.08**

Table 3. Results on the hallucination subset of MME. The best performances within each setting are **bolded**.

IMD	CMPC	POPE Accuracy	F1 Score	CHAIR $CHAIR_i$	$CHAIR_s$
		79.98	79.34	55.6	17.8
✓		85.10	84.68	48.2	13.6
	✓	82.31	81.86	54.2	16.2
✓	✓	86.04	85.70	47.0	12.7

Table 4. Ablation of the core components on POPE Popular and CHAIR for LLaVA 1.5.

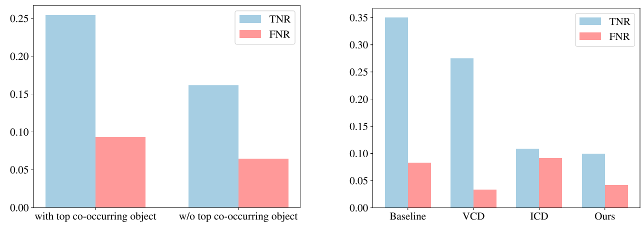
generalization in different kinds of tasks.

Distortion	POPE Accuracy	F1 Score	CHAIR $CHAIR_i$	$CHAIR_s$
Attention mask	84.92	85.21	48.4	13.9
Value noise addition	85.72	85.60	49.5	13.8
Value mask	86.04	85.70	47.0	12.7

Table 5. Ablation of different kinds of distortion used in our method.

Analysis of hallucination with statistically significant cross-modal object existence. To assess the effectiveness of our method in mitigating hallucinations caused by spurious inter-modality correlations, we evaluate the hallucination rates of LLaVA 1.5 on the MS-COCO and A-OKVQA subsets of the POPE dataset in scenarios with statistically significant cross-modal object existence, which often induces spurious correlations. Figure 5(a) illustrates that the statistically significant co-existence leads to a serious hallucination both on the true negative rate and the false negative rate. This finding indicates that spurious inter-modality correlations can cause both the neglect of relevant regions of interest and an over-reliance on irrelevant regions. Figure 5(b) demonstrates that our method effectively reduces both true negative and false negative hallucinations stemming from co-existence, underscoring its ability to mitigate hallucinations from spurious inter-modality correlations.

Analysis of hallucination with position bias. We investigate the distribution of hallucinations in LLaVA-1.5 with respect to token position. It can be observed from Table.6 that the hallucinations occur more frequently in the head portion of the image tokens, a phenomenon caused by the inherent limitations of the RoPE (Rotary Position Embedding) mechanism. By contrast, with the integration of our Cross-Modal Position Calibration (CMPC), the model not only exhibits a significant reduction in hallucinations but also achieves a more even distribution of hallucinations across different parts of image tokens. These findings highlight the effectiveness of our CMPC strategy in mitigating position bias and reducing hallucinations in multimodal models.



(a) (b)

Figure 5. The comparison of the hallucination rate of LLaVA 1.5 on the POPE dataset. 'TNR' and 'FNR' denote the true negative rate and the false negative rate of VQA, respectively. (a) The hallucination rate of the existence of objects with and without the co-existence with their top co-occurring object in the image. (b) The hallucination rate of the object's existence for different decoding methods with their top co-occurring object. Concretely, we estimate the mean hallucination rate on 5 pairs of objects with a high object existence rate.

Method	Head Part	Middle Part	Tail Part
Baseline	28%	21%	15%
Ours	11%	13%	10%

Table 6. The effect of our method in mitigating the positional bias of the hallucination occurrence frequency.

5. Conclusion

In this paper, we have introduced CMAC, a novel framework addressing the hallucination problem of LVLMs with contrastive decoding. In this approach, we propose the Inter-Modality Decoding which estimates a distorted distribution favored hallucination by selectively masking the value vectors associated with high cross-modal attention weights, alleviating both the spurious inter-modality correlation and the uni-modal over-reliance. Furthermore, a Cross-Modal Position Calibration module is introduced to alleviate the overlook of essential visual content caused by position bias. Extensive experiments conducted on diverse benchmarks and LVLMs confirm the efficacy of the proposed method.

Acknowledgments

This work is supported in part by the National Key R&D Program of China (NO. 2024YFB3908503 and 2024YFB3908500), and in part by the National Natural Science Foundation of China (NO. 62322608 and NO. 62495081).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[2] Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Ping Chen, Xiaoqin Zhang, and Shijian Lu. Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29915–29926, 2025. 3

[3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 1

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 1, 2, 6

[5] Junzhe Chen, Tianshu Zhang, Shiyu Huang, Yuwei Niu, Linfeng Zhang, Lijie Wen, and Xuming Hu. Ict: Image-object cross-level trusted intervention for mitigating object hallucination in large vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4209–4221, 2025. 3

[6] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 1, 2

[7] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.

[8] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 1, 2

[9] Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. Halc: Object hallucination reduction via adaptive focal-contrast decoding. *arXiv preprint arXiv:2403.00425*, 2024. 3, 4

[10] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 6

[11] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023. 1

[12] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: towards general-purpose vision-language models with instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2024. Curran Associates Inc. 2, 6

[13] Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14303–14312, 2024. 1, 3

[14] Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18135–18143, 2024. 3

[15] Junjia Huang, Pengxiang Yan, Jinhang Cai, Jiyang Liu, Zhao Wang, Yitong Wang, Xinglong Wu, and Guanbin Li. Dreamlayer: Simultaneous multi-layer generation via diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3357–3366, 2025. 1

[16] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 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, 2024. 3, 6

[17] Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-introspective decoding: Alleviating hallucinations for large vision-language models. *arXiv preprint arXiv:2408.02032*, 2024. 3, 4

[18] Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27036–27046, 2024. 3

[19] Minchan Kim, Minyeong Kim, Junik Bae, Suhwan Choi, Sungkyung Kim, and Buru Chang. Exploiting semantic reconstruction to mitigate hallucinations in vision-language models. In *European Conference on Computer Vision*, pages 236–252. Springer, 2025. 3

[20] Sicong Leng, Yun Xing, Zesen Cheng, Yang Zhou, Hang Zhang, Xin Li, Deli Zhao, Shijian Lu, Chunyan Miao, and Lidong Bing. The curse of multi-modalities: Evaluating hallucinations of large multimodal models across language, visual, and audio. *arXiv preprint arXiv:2410.12787*, 2024. 2, 5

[21] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882, 2024. 1, 3, 4, 6, 7

[22] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2

[23] Jiaming Li, Jiacheng Zhang, Jichang Li, Ge Li, Si Liu, Liang Lin, and Guanbin Li. Learning background prompts to discover implicit knowledge for open vocabulary object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16678–16687, 2024. 1

[24] Jiaming Li, Zhijia Liang, Weikai Chen, Lin Ma, and Guanbin Li. Guided: Granular understanding via identification, detection, and discrimination for fine-grained open-vocabulary object detection. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 1

[25] Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. Silkie: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665*, 2023. 3

[26] Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. *arXiv preprint arXiv:2210.15097*, 2022. 3

[27] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 5

[28] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2023. 3

[29] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 1, 2, 6

[30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1, 2

[31] Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. *arXiv preprint arXiv:2407.21771*, 2024. 3, 6

[32] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018. 5

[33] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023. 3

[34] Wei Suo, Lijun Zhang, Mengyang Sun, Lin Yuanbo Wu, Peng Wang, and Yanning Zhang. Octopus: Alleviating hallucination via dynamic contrastive decoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29904–29914, 2025. 3

[35] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1

[36] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 1

[37] Kuo Wang, Quanlong Zheng, Junlin Xie, Yanhao Zhang, Jinguo Luo, Haonan Lu, Liang Lin, Fan Zhou, and Guanbin Li. Free-moref: Instantly multiplexing context perception capabilities of video-mllms within single inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22499–22508, 2025. 1

[38] Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. *arXiv preprint arXiv:2403.18715*, 2024. 3, 4, 5, 6

[39] Sangmin Woo, Jaehyuk Jang, Donguk Kim, Yubin Choi, and Changick Kim. Ritual: Random image transformations as a universal anti-hallucination lever in lvlms. *arXiv preprint arXiv:2405.17821*, 2024. 3

[40] Sangmin Woo, Donguk Kim, Jaehyuk Jang, Yubin Choi, and Changick Kim. Don't miss the forest for the trees: Attentional vision calibration for large vision language models. *arXiv preprint arXiv:2405.17820*, 2024. 1, 3, 4

[41] Junfei Wu, Qiang Liu, Ding Wang, Jinghao Zhang, Shu Wu, Liang Wang, and Tieniu Tan. Logical closed loop: Uncovering object hallucinations in large vision-language models. *arXiv preprint arXiv:2402.11622*, 2024. 3

[42] Zhihe Yang, Xufang Luo, Dongqi Han, Yunjian Xu, and Dongsheng Li. Mitigating hallucinations in large vision-language models via dpo: On-policy data hold the key. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10610–10620, 2025. 3

[43] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 2

[44] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13040–13051, 2024. 2

[45] Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong

Chen. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*, 2023. 3

[46] Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12944–12953, 2024. 3

[47] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 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, 2024. 3

[48] Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*, 2024. 3

[49] Zihao Yue, Liang Zhang, and Qin Jin. Less is more: Mitigating multimodal hallucination from an eos decision perspective. *arXiv preprint arXiv:2402.14545*, 2024. 3

[50] Bohan Zhai, Shijia Yang, Xiangchen Zhao, Chenfeng Xu, Sheng Shen, Dongdi Zhao, Kurt Keutzer, Manling Li, Tan Yan, and Xiangjun Fan. Halle-switch: Rethinking and controlling object existence hallucinations in large vision language models for detailed caption. *arXiv preprint arXiv:2310.01779*, 2023. 3

[51] Yue Zhang, Leyang Cui, Wei Bi, and Shuming Shi. Alleviating hallucinations of large language models through induced hallucinations. *arXiv preprint arXiv:2312.15710*, 2023. 4

[52] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*, 2023. 3

[53] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 2