IBD: Alleviating Hallucinations in Large Vision-Language Models via Image-Biased Decoding

Lanyun Zhu[1] Deyi Ji[2] Tianrun Chen[3] Peng Xu[2] Jieping Ye[2] Jun Liu[4]
[1]Singapore University of Technology and Design [2]Alibaba [3]Zhejiang University [4]Lancaster University

Abstract

Despite achieving rapid developments and with widespread applications, Large Vision-Language Models (LVLMs) confront a serious challenge of being prone to generating hallucinations. An over-reliance on linguistic priors has been identified as a key factor leading to these hallucinations. In this paper, we propose to alleviate this problem by introducing a novel image-biased decoding (IBD) technique. Our method derives the next-token probability distribution by contrasting predictions from a conventional LVLM with those of an image-biased LVLM, thereby amplifying the correct information highly correlated with image content while mitigating the hallucinatory errors caused by excessive dependence on text. We further conduct a comprehensive statistical analysis to validate the reliability of our method, and design an adaptive adjustment strategy to achieve robust and flexible handling under varying conditions. Experimental results across multiple evaluation metrics verify that our method, despite not requiring additional training data and only with a minimal increase in model parameters, can significantly reduce hallucinations in LVLMs and enhance the truthfulness of the generated response.

1. Introduction

The rapid development of large language models (LLMs) [4, 29] in recent years marks a significant step toward achieving artificial general intelligence (AGI). Large vision-language models (LVLMs) [24, 35–37] have further extended the capabilities of LLMs to the visual domain, demonstrating impressive performance across various tasks such as image captioning [20, 34], grounding [26, 40] and segmentation [46]. Despite their notable successes, LVLMs still encounter numerous challenges that impede their real-world utility. Among these challenges, the phenomenon of hallucinations [13, 25] is particularly critical, with its frequent occurrence prompting extensive worries about the safety and stability of artificial intelligence systems.

In the field of vision-language models, the term 'hallucination' denotes the generation of information by the model

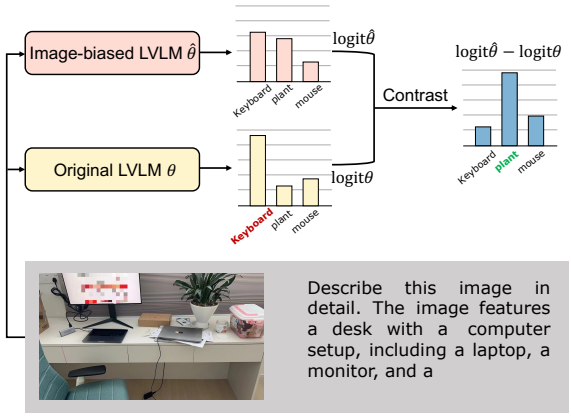


Figure 1. An illustrative example of our method. Texts highlighted in red and green indicate erroneous prediction and correct prediction generated by the original LVLM and contrastive results, respectively.

that is either unrelated or incorrect with respect to the given text and image inputs. Recent studies [11] suggest that the standard pretraining mechanism of LLMs, which targets the maximum-likelihood-based next token prediction, is a significant contributor to the occurrence of hallucinations. This mechanism may cause models to depend excessively on superficial patterns found in the training data rather than developing a substantial understanding of fundamental logical principles [15]. Specifically for the vision-language models, it has been observed that LVLMs fine-tuned from language models tend to depend excessively on linguistic priors during the autoregressive text generation process [10, 19, 22]. As a result, responses are often simply continued based on the text generated from the previous time steps, neglecting to draw logical inferences from the actual visual content within the input image. This issue of language over-reliance is particularly severe in the generation of long text, with empirical observations [44] indicating that the later segments of the generated response are more susceptible to hallucinations.

In this work, we aim to address the issue of hallucinations in LVLMs by proposing a novel contrastive decod-

ing technique [21] that is designed based on the underlying cause of these hallucinations to extract accurate information. Our proposed method involves computing a more reliable next-token probability distribution by contrasting the predictions of the original model with those of an image-biased model that focuses more the image information. The image-biased model is created by modifying the attention weight matrix structure within the original model without altering its parameters. This approach emphasizes the knowledge of the image-biased model and diminishes that of the original model, which may be text-biased, thus encouraging the extraction of correct content while suppressing hallucinations resulting from textual over-reliance. An illustrative example of our method is presented in Fig.1. The original model erroneously predicts 'keyboard' as the next token, a term often associated with 'computer' in linguistic priors but absent in the input image. Upon using the image-biased model, the correct token 'plant' receives a significantly higher probability boost. Consequently, the contrast between these two predictions can serve as an effective indicator to select the accurate tokens and mitigate hallucinations.

To ensure the reliability of our method, we conduct a statistical analysis to verify the alignment between the tokens receiving the highest probability boost and the correct tokens devoid of hallucinations. The results reveal a significant correlation in certain conditions, demonstrating the credibility of our method. Furthermore, we identify potential scenarios where our method may fail and, in response, develop a dynamic adjustment mechanism to address such issues. By incorporating these designs, our method, named **I**mage-**B**iased **D**ecoding (IBD), demonstrates notable effectiveness in mitigating hallucinations. The benefits of IBD can be summarized in three key aspects: (1) **Minimal Overhead**. IBD operates with low extra parameter and data costs, distinguishing it from previous methods like [23, 41], which necessitate extra training data, or [39, 44] which demand a considerable expansion of model parameters. (2) **Comprehensive Processing Capability**. Unlike earlier contrastive decoding strategies [19, 21], IBD is designed based on an in-depth statistical analysis that takes into account the unique properties of different vocabulary types, allowing it to adaptively handle a diverse range of situations flexibly. (3) **Superior Performance**. Extensive validation experiments are conducted to benchmark IBD against other advanced methods. The leading results across multiple metrics demonstrate the high effectiveness of IBD.

2. Related Work

Large Vision-Language Models. The introduction of large language models (LLMs) [4, 29] has marked the advent of a new era in artificial intelligence. Initially, LLMs were limited to text processing. Recent advancements have expanded their capabilities to process images, leading to the development of Large Vision-Language Models (LVLMs) [1, 8, 24, 45] in a multi-modal manner. These LVLMs have achieved impressive and generalizable performance across multiple tasks, such as image captioning [20], visual question answering [42], object detection [38] and image segmentation [18]. Despite their successes, similar to LLMs, LVLMs are prone to generating hallucinations [22], which significantly hampers their robust application. Our research aims to address the issue of hallucinations in LVLMs to improve their practical utility and reliability.

Hallucinations in VLMs and LVLMs. Hallucination refers to the problem that a model generates information that is either irrelevant or incorrect with respect to the given context. This issue has attracted increased attention within the domain of VLMs [16, 27]. Recent researchers [2, 32, 43] have started to focus intensively on the hallucination phenomenon in LVLMs. Some approaches [23, 30, 41] attempt to mitigate hallucinations by fine-tuning the LVLMs on fine-grained datasets, while others [33, 39, 44] propose to use supplementary networks to identify hallucinated objects or to rewrite the output text from LVLMs. However, these techniques typically necessitate a substantial number of additional model parameters or training data. Our method differs from these approaches by focusing only on refining the model's decoding process, thereby effectively reducing hallucinations without requiring too many additional parameters or data.

Decoding Method. Decoding method determines the generation of text tokens at each time step within language models. Traditional decoding strategies such as beam search [3], top-k decoding [9], and sampling methods [11], despite their widespread use, are prone to producing hallucinatory content. Recent research [6, 7, 14, 19, 21, 31] has made attempts to address this issue by proposing better decoding methods. For instance, [19] uses contrastive decoding in LVLMs; however, this technique relies on randomly generated noise as input, which can introduce uncontrollable and unstable informational perturbations. In contrast to these existing approaches, our method delves into the fundamental causes of hallucinations in LVLMs and tailors a comprehensive method in consideration of varying textual words, leading to an efficient and robust solution for handling a wide array of scenarios. Table.1 presents the comparison results, where our approach demonstrates superior performance than other decoding methods.

3. Method

Recent Large Vision-Language Models (LVLMs) typically utilize a visual image v and a text t as inputs to generate a text response. The process begins with passing the input image v through a visual encoder, followed by a feature space projector to generate a set of visual tokens. These

visual tokens are subsequently concatenated with the input text tokens to serve as the input for the LVLM's language model. Within the framework, the LVLM generates the response $y = \{y_1, y_2, ..., y_{N_y}\}$ with N_y time steps in an auto-regressive manner, with the probability distribution of generating token y_i at the i-th time step being represented as $p_\theta\left(y_i|v, t, y_{<i}\right) = \text{Softmax}\left(\text{logit}_\theta\left(y_i|v, t, y_{<i}\right)\right)$, where θ refers to the LVLM model, and $y_{<i} = \{y_1, y_2, ..., y_{i-1}\}$ denotes the response generated from the previous time steps.

In contrast to vanilla LVLMs, our method does not directly utilize the probability distribution obtained through the aforementioned process for prediction. Instead, to alleviate hallucinations, we obtain next-token prediction by contrasting the ordinary prediction with an image-biased prediction. Specifically, alongside the existing model θ, we introduce a new model $\hat{\theta}$. Different from θ, which theoretically balances the use of both image v and text $\{t, y_{<i}\}$ for prediction, $\hat{\theta}$ exhibits a higher dependency and places greater emphasis on the information contained within the image v. With the prediction logits $\text{logit}_{\hat{\theta}}\left(y_i|v, t, y_{<i}\right)$ obtained using $\hat{\theta}$, we compute the next-token-probability by:

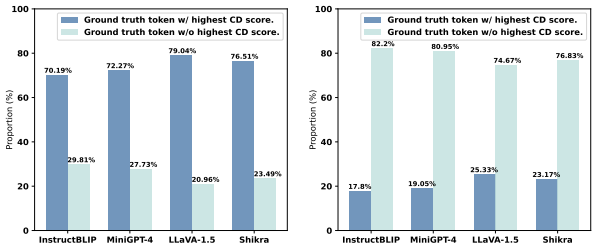
$$p\left(y_i|y_{<i}\right) = \text{Softmax}\left(\mathcal{L}_{CD}\right), \tag{1}$$

$$\mathcal{L}_{CD} = \text{logit}_{\hat{\theta}}\left(y_i|v, t, y_{<i}\right) - \text{logit}_\theta\left(y_i|v, t, y_{<i}\right), \tag{2}$$

Where $\mathcal{L}_{CD}$ is named as CD score. The motivation for introducing $\hat{\theta}$ and the contrastive decoding method is based on the hallucination characteristics observed in LVLMs. Previous studies [10, 22] have identified the excessive dependence on the language model's linguistic priors as a critical contributor to LVLM's hallucinations. In particular, it has been observed that LVLMs can occasionally disregard input image's information but treat the next-token prediction as a purely text-based continuation task. Such an approach can lead to the generation of responses that are unrelated to the input image and are therefore text-biased hallucinations. Based on these observations, we propose a hypothesis that, using the model $\hat{\theta}$ that focuses more on input image information when making predictions in each time step, the token with the highest probability boost from θ to $\hat{\theta}$ among the candidate pool is more likely to be the correct token without text-biased hallucinations. We provide a detailed description of model $\hat{\theta}$ in Sec.3.1; and in Sec.3.2, we examine our hypothesis regarding the differences in results when utilizing θ and $\hat{\theta}$.

3.1. Image-biased Model

We employ a simple yet effective approach to construct the image-biased model $\hat{\theta}$, by merely adjusting the attention weight matrices within the vanilla model θ. Within each attention weight matrix of LVLM's language model, we apply an amplification coefficient to the image tokens, thereby increasing the attention's focus on visual information while



(a) Content Words (b) Function Words

Figure 2. (a) and (b) respectively present the statistical results for content words and function words in the COCO Caption dataset. **Dark** bars represent the proportion of ground tokens with the highest CD score among all candidate tokens, while light bars represent the proportion of ground tokens without the highest CD score. The statistical results on four LVLMs are reported, including InstructBLIP, MiniGPT-4, LLaVA-1.5 and Shikra.

minimizing its interaction with textual features. This amplification coefficient c is a pre-defined numerical value, which is conditionally added to the attention's query-key multiplication result when the key corresponds to an image token. To be specific, in the i-th time step for generating the text response, we denote the tokens for the l-th layer of the LVLM as $\mathbf{T}^l = \{\mathbf{T}^l_v, \mathbf{T}^l_o\}$, where $\mathbf{T}^l_v = \{\mathbf{T}^l_{v,k}\}_{k=1}^{N_v}$ and $\{\mathbf{T}^l_{o,k}\}_{k=1}^{N_o}$ refer to the tokens of the input image v and other components (input text, previously generated response, padding tokens), respectively. After that, we use the following formulation to compute each item $W^l_{m,n}$ of the attention's weight matrix W^l:

$$W^l_{m,n} = \text{Softmax}\left(\frac{Q^l_m (K^l_n)^{\mathrm{T}}}{\sqrt{D}} + c_{m,n} + M\right), \tag{3}$$
$$c_{m,n} = \epsilon \quad \text{if} \quad K^l_n \leftarrow \mathbf{T}^l_v \quad \text{else} \quad 0,$$

where Q^l and K^l refer to the query and key tokens derived from $\mathbf{T}^l$, M represents the casual mask, and D denotes the feature dimension. $K^l_n \leftarrow \mathbf{T}^l_v$ represents that the n-the key token K^l_n corresponds to an image token. Utilizing the method, we encourage the attention mechanism to place greater focus on image information, consequently making the constructed model $\hat{\theta}$ to become image-biased. Note that $\hat{\theta}$ shares the same weight parameters with θ, thus preventing the additional memory usage to load other parameters during the inference stage. In Sec.4.3, we will discuss the detailed settings for the hyperparameter ϵ in Eq.3.

3.2. Hypothesis Testing

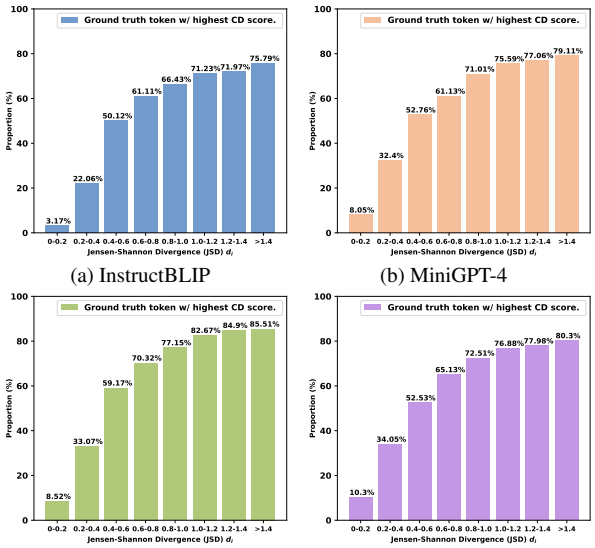
Based on $\hat{\theta}$, we evaluate the hypothesis presented above, i.e., in each time step, candidate token with the highest probability boost from θ to $\hat{\theta}$ (CD score $\mathcal{L}_{CD}$ as Eq.2) is more likely to be the correct response without text-biased hallucinations. We conduct experiments on COCO datasets to test this hypothesis, and we find the following two patterns:

Difference for Content Words and Function Words. Our first finding is that the proposed hypothesis holds true for content words but does not apply to function words. In linguistics, content words refer to words including rich information like most of the nouns and adjectives; function words, on the other hand, refer to other words for forming sentence structures but with less informational content, such as 'a', 'the', 'to', 'and', etc. We employ ChatGPT to categorize each word in the COCO captions into these two types. For each time step of a caption text, we compute the CD scores $\mathcal{L}_{CD}$ using Eq.2, and assess whether the ground truth token attains the highest CD score among all candidate tokens[1]. This assessment is conducted across all caption words with the ground truth classified as either content or function words, followed by a statistical analysis shown in Fig.2. Our findings indicate that, for content words, the ground truth tokens frequently receive the highest scores, thus validating our hypothesis for these terms. In contrast, for function words, the hypothesis is not supported, and the ground truth tokens frequently score lower than the other hallucinatory candidate tokens. We attribute this pattern to the LVLM's reliance on different types of information when predicting different words. Content words, often closely associated with image content, benefit from a model's increased focus on image data. Conversely, function words, devoid of visual information and heavily tied to grammatical constructs, depend more on the linguistic priors within the LLM. Consequently, using an image-biased $\hat{\theta}$ may increase the probability of predicting content tokens while reducing the probability of predicting function tokens, thus leading to higher CD scores for the content ground truth tokens and lower scores for the function ground truth tokens.

Impact of Prediction Similarity between θ and $\hat{\theta}$. Another interesting finding is that the proposed hypothesis may not hold when the predictions $p_\theta\left(y_i|v,t,y_{<i}\right)$ from θ and $p_{\hat{\theta}}\left(y_i|v,t,y_{<i}\right)$ from $\hat{\theta}$ are too similar. Specifically, we use d_i, calculated as the Jensen-Shannon divergence (JSD), to quantify the similarity of the predictions by $d_i = \text{JSD}\left(p_\theta\left(y_i|v,t,y_{<i}\right)||p_{\hat{\theta}}\left(y_i|v,t,y_{<i}\right)\right)$. Our statistical results, as shown in Fig.3, suggest that a reduction in d_i correlates with a greater probability that the candidate token with the highest CD score is NOT the actual ground truth in the i-th time step. When the prediction similarity between θ and $\hat{\theta}$ is too high, the CD score, computed as the prediction difference, would be low in value and likely to be influenced by noise, thus leading to a reduced accuracy by using it for identifying the correct token. This finding is consistent with the observations reported in the previous research [7, 21].

A low value of d_i could be attributed to two reasons: (1) the prediction from θ for y_i might already be heavily image-dependent, and thus, further enhancing image focus in $\hat{\theta}$ does not lead to a significant change in the prediction outcome; (2) the model's hallucinations by over-relying on text coincidentally aligns with the actual content within the image, so the model is capable of achieving similar predictions for y_i when either over-relying on textual information in θ or image in $\hat{\theta}$. Note that we refer to reason (2) as 'benign text-biased hallucination. In fact, this phenomena is common in LVLMs, as the LVLM's text-biased hallucinations tend to be biased toward the learned associations from the training set [44]. These associations may also likely be present in the test images. For instance, upon predicting the word 'computer', an LVLM might frequently predict 'keyboard' subsequently without relying on image information, because 'computer' and 'keyboard' are commonly seen together in the training corpus. If a keyboard is indeed in the test image that also contains a computer, this sort of benign text-biased hallucinations will not lead to an erroneous prediction.

[1]Candidate tokens include one ground truth token and the other hallucinatory candidate tokens



(a) InstructBLIP (b) MiniGPT-4 (c) LLaVA-1.5 (d) Shikra

Figure 3. Statistical results to illustrate the relationship between the prediction similarity of θ and $\hat{\theta}$ and the proportion of ground tokens having the maximum CD score among candidate tokens. X-axis denotes the range of Jensen-Shannon divergence (JSD) d_i between the prediction results from θ and $\hat{\theta}$. d_i is scaled by 1.5×10^4. A higher d_i indicates lower similarity. Y-axis represents, for all time steps with its d_i falling into each range, the proportion of time steps where the ground truth token has the highest CD score among all candidate tokens.

3.3. Dynamic Adjustment For Decoding

The above analysis reveals that an over-reliance on textual information is not always detrimental. In the case of predicting function words, a dependence on language priors is appropriate, whereas a bias toward images is detrimental. Moreover, when predictions from θ and $\hat{\theta}$ are closely

aligned, there is no substantial difference between text-reliant and image-reliant outcomes. Consequently, it is impractical to employ CD scores for the prediction of such terms. Through intuitive analysis, we assert that even when using the vanilla maximum likelihood decoding (greedy decoding) method, these terms are less prone to suffer from errors caused by text-biased hallucinations. This is because function words are generally easy to predict. While terms with similar predictions from θ and $\hat{\theta}$—either already emphasizing image information when predicting or exhibiting benign text-biased hallucinations—are less likely to be the erroneous information caused by the over-reliance on text.

Building upon the above discussion, we introduce a dynamic mechanism for token prediction, which adaptively adjusts the balance between the traditional maximum likelihood decoding method and the proposed image-biased contrastive decoding. This self-adaptive adjustment is implemented according to the state of the current decoding step. Formally,

$$\begin{aligned} y_i \sim \text{Softmax}\left(\text{logit}_\theta\left(y_i|v,t,y_{<i}\right) + \alpha \cdot I \cdot \mathcal{L}_{CD}\right), \\ \text{where } I = \text{Min}\{I_{sim}, I_{con}\}, \end{aligned} \tag{4}$$

where α is a scaling factor, I_{sim} and I_{con} are two indicators. I_{sim} reflects the difference between the predictions from θ and $\hat{\theta}$, and I_{con} reflects the probability that the token in the current decoding time step is a content word. In practice, we compute I_{sim} as the Jensen-Shannon divergence by:

$$I_{sim} = \text{JSD}\left(p_\theta\left(y_i|v,t,y_{<i}\right) || p_{\hat{\theta}}\left(y_i|v,t,y_{<i}\right)\right). \tag{5}$$

To calculate I_{con}, one might consider employing an existing POS tagging tool to distinguish whether a predicted token is a content or function word. This method, however, is prone to semantic ambiguities due to subword tokenization. For instance, the token 'on' generated in a time step might be part of the content word 'onion'. Directly labeling 'on' as a function word could therefore introduce inaccuracies. To address this issue, we instead adopt an implicit approach. Prior research [7] has shown that when employing an early exit strategy [28] in language models to predict function words, the LLM tends to determine the token to generate within the middle layers, and keeping the predictions almost unchanged in the subsequent higher layers. In contrast, when predicting content words that are more complex, the LLM continues to change its predictions in the last few layers. Levering these findings, we employ the same early exiting method as [7] to get a prediction $\widetilde{p}_\theta\left(y_i|v,t,y_{<i}\right)$ from the LVLM's middle layer (the 24-th layer in our setting). We then define I_{con} as the distance between the predictions from the final layer and those from the intermediate layer, which is formally calculated as:

$$I_{con} = \text{JSD}\left(p_\theta\left(y_i|v,t,y_{<i}\right) || \widetilde{p}_\theta\left(y_i|v,t,y_{<i}\right)\right). \tag{6}$$

I_{sim} and I_{con} are used in Eq.4 to complete the prediction process. This dynamic approach allows the model to rely more on maximum likelihood when predicting function words and words that elicit similar predictions from θ and $\hat{\theta}$, thereby mitigating the negative impact that may arise from a reduced ability of the CD score to accurately identify the correct tokens in certain instances. This harmonization of multiple strategies ensures that the model's performance is not unduly hindered by potential weakness in any single approach, particularly when dealing with some specific types of words as we have discussed before.

3.4. Full Method

Building upon the approaches illustrated above, we further advance our method by introducing two minor yet effective improvements. The first improvement is to finetune the image-biased model $\hat{\theta}$. In our method, $\hat{\theta}$ is derived by modifying the attention weight matrices without altering the parameter weights. However, this modification on model structures might introduce noise, which in turn could impact the reliability of the CD score. To mitigate the issue, we finetune $\hat{\theta}$ on the COCO caption dataset to adapt it better to the modified attention architecture. Note that we do not apply the frequently-used LoRA for finetuning, which requires an excessive number of additional parameters. Instead, we utilize the more lightweight prompt tuning technique by adding a small set of learnable prompts P to the LLM's inputs. With P obtained through finetuning, the output of $\hat{\theta}$ at the i-th time step can be formulated as $\text{logit}_{\hat{\theta}}\left(y_i|P,v,t,y_{<i}\right)$. The second improvement draws inspiration from previous methods [21] by employing an adaptive plausibility constraint, which is realized by selectively considering only a portion of tokens that possess high enough output probabilities as the candidates for prediction. This approach helps to prevent the frequent occurrence of false positives and false negatives in contrastive decoding methods [7, 21]. Incorporating these improvements, the full method to predict a token in a time step i is written as:

$$\begin{aligned} y_i \sim \text{Softmax}\left(\text{logit}_\theta\left(y_i|v,t,y_{<i}\right) + \alpha \cdot I \cdot \mathcal{L}_{CD}\right), \\ \text{subject to } y_i \in \mathcal{V}_{head}\left(y_{<i}\right), \end{aligned} \tag{7}$$

$$\begin{aligned} \text{where } \mathcal{L}_{CD} &= \text{logit}_{\hat{\theta}}\left(y_i|P,v,t,y_{<i}\right) - \text{logit}_\theta\left(y_i|v,t,y_{<i}\right), \\ I &= \text{Min}\{I_{sim}\,(\text{Eq.5}),\ I_{con}\,(\text{Eq.6})\}, \\ \mathcal{V}_{\text{head}}\left(y_{<i}\right) &= \{y_i \in \mathcal{V}: \\ &\qquad p_\theta\left(y_i|v,t,y_{<i}\right) \geq \beta \max_w p_\theta\left(w|v,t,y_{<i}\right)\}, \end{aligned} \tag{8}$$

where $\mathcal{V}$ is LVLM's output vocabulary and β is a hyperparameter between 0 to 1.

Figure 4. An example to show the problem of image-biased hallucinations in LVLMs. Texts highlighted in red and green indicate erroneous information and correct information generated by LLaVA-1.5, respectively.

3.5. Discussion: Are there Image-biased Hallucinations?

This study builds on the existing understanding that hallucinations in LVLMs typically stem from an over-reliance on textual information. A question that naturally arises is whether LVLM's hallucinations exclusively arise from text dependency, and if some can also be attributed to the excessive reliance on input image. An example shown in Fig.4 demonstrates that such image-biased hallucinations indeed exist. In the example, a photo showing the renowned soccer star Cristiano Ronaldo playing table tennis is presented to an LVLM with the question about whether the man is a professional table tennis player. The LVLM responses incorrectly. Subsequently, the image is cropped to exclude the table tennis racket, and the modified image is resubmitted to the LVLM, which in turn provides the correct response. This result suggests that the LVLM's attention to the paddle region impairs its judgment, thereby validating the existence of image-biased hallucinations.

Our empirical observation suggests that such image-biased hallucinations frequently occur when there is an inconsistency between the visual content and the language model's world knowledge. For example, as shown in Fig.4, determining if the man in the image is a professional table tennis player relies on the language model's knowledge that Cristiano Ronaldo is a footballer, a fact at odds with the presence of a table tennis racket for another sport in this image. Consequently, overreliance on the visual cue of the racket, without proper integration of the model's world knowledge, may lead to an incorrect inference that the man is a table tennis professional. Such misjudgments, however, are uncommon in the mainstream evaluation frameworks for LVLMs. These frameworks typically encompass tasks such as image captioning, which necessitates a limited scope of world knowledge and focus more on the direct descriptions of visual content, and question-answering tasks that typically process straightforward questions without excessive interference. In these cases, the visual content is less prone to conflicting with the language model's world knowledge, thereby making image-biased hallucinations a rarer phenomenon.

Given the insights obtained from our analyses, and taking into account the characteristics of the existing evaluation frameworks, we have chosen not to create targeted solutions for the image-biased hallucinations. Instead, we propose the development of a more comprehensive assessment method for image-biased hallucinations and the corresponding solutions as a topic for future research.

Table 1. Evaluation results on CHAIR metric. Smaller values indicate fewer hallucinations.

	InstructBLIP		MiniGPT-4		LLaVA-1.5		Shikra	
Method	C_S	C_I	C_S	C_I	C_S	C_I	C_S	C_I
Greedy	30.0	14.5	24.2	8.2	20.6	6.2	22.0	7.0
Nucleus	30.4	15.7	23.6	8.3	26.2	8.5	22.6	7.6
Beam Search	21.4	7.2	23.6	7.8	18.8	5.9	20.2	6.4
ReCaption	17.2	6.6	21.8	8.1	13.8	5.0	13.9	5.5
Woodpecker	15.5	6.5	21.5	7.8	13.4	4.8	13.6	**5.1**
CD	22.0	7.1	23.8	8.3	20.9	6.0	20.4	6.4
DoLa	22.2	7.1	24.2	8.2	20.4	6.3	20.2	6.3
VCD	19.7	7.0	23.7	8.0	18.9	5.7	19.5	6.2
OPERA	16.6	6.8	22.6	8.2	14.2	5.2	14.2	5.9
IBD	**15.0**	**6.2**	**21.0**	**7.4**	**12.7**	**4.5**	**13.2**	5.2

4. Experiments

4.1. Settings

Implementation Details. We set ϵ in Eq.3 to 2, α in Eq.4 and Eq.7 to 1.5×10^4, and β in Eq.8 to 0.1.$\widetilde{p}_\theta\left(y_i|v,t,y_{<i}\right)$ in Eq.6 for computing I_{con} is derived from LVLM's 24-th layer.

Baseline Models. Following previous research [14], we conduct performance validation and comparison across four mainstream large vision-language models, including InstructBLIP [8], MiniGPT-4 [45], LLaVA-1.5 [24] and Shikra [5]. All these models utilize an LLM with 7 billion parameters.

4.2. Main Results

Following [14], we test our approach using three widely-used metrics, including CHAIR evaluation, GPT-4 assisted evaluation and GPT-4V assisted evaluation.

CHAIR Evaluation. Caption Hallucination Assessment with Image Relevance (CHAIR) is a commonly-used metric for evaluating the degree of object hallucination in image captioning tasks. It assesses the proportion of objects that are present in the generated captions but absent in the

Table 2. Comparison with other assistance-based methods in terms of additional parameters/tools and the CHAIR metric. Lower values of C_S and C_I indicate fewer hallucinations.

Method	Additional Param/Tool	MiniGPT-4		LLaVA-1.5	
		C_S	C_I	C_S	C_I
ReCaption	GPT4	21.8	8.1	13.8	5.0
WoodPecker	273M	21.5	7.8	13.4	4.8
IBD	**68K**	**21.0**	**7.4**	**12.7**	**4.5**

ground truth. CHAIR encompasses two dimensions, denoted as C_S and C_I, which evaluate the degree of object hallucination from sentence-level and image-level perspectives, respectively. C_S and C_I are computed as:

$$C_S = \frac{|\{\text{captions w/ hallucinated objects}\}|}{|\{\text{all captions}\}|}, C_I = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all mentioned objects}\}|}. \tag{9}$$

We follow the same setting as [14] by employing the validation set of the COCO 2014 dataset for our evaluation, and using `Please describe this image in detail.` as LVLM's text input. We present the comparison results in terms of C_S and C_I in Table.1. The methods compared in the table are categorized into three types: (1) the traditional decoding methods, including Greedy Decoding, Nucleus decoding [11] and Beam Search decoding [3]; (2) decoding methods specifically designed to address hallucinations, including vanilla CD (CD) [21], visual CD (VCD) [19], DOLA [7] and OPERA [14]; and (3) assistance-based methods that employ an additional model to mitigate hallucinations or rewrite descriptions, including ReCaption [33] and Woodpecker [39]. The results show that our IBD can consistently outperform all other types of methods across different baseline models, demonstrating its high effectiveness. Also note that IBD requires only a minimal number (74K) of additional parameters for the prompt P beyond the baseline model. This distinguishes it from other assistance-based methods, which need to use additional LLMs or network architectures with an excessive number of additional parameters. As shown in Table.2, compared to these methods, IBD is both more effective and efficient.

GPT-4 Assisted Evaluation. A limitation of CHAIR is that it can only assess object-existence-level hallucination. To further measure hallucinations at other levels, such as attributes and relations, we follow the method in [14] by employing GPT-4 to evaluate the discrepancies between the descriptions generated by the LVLM and the ground truth descriptions on the VG dataset [17], thus assessing the degree of hallucinations in a more comprehensive manner. For a fair comparison, we prompt GPT-4 using the same method as [14], asking GPT4 to produce the assessment for each description with six metrics. Readers can refer to Sec.4.4 of [14] for detailed explanations of each metric. We conduct comparisons on four metrics that are highly correlated to the hallucination: the number of hallucinated sentences per image (HSPI), the number of hallucinated words per image (HWPI), the ratio of hallucinated sentences (HSR), and the ratio of hallucinated words (HWR). The results of these comparisons are presented in Fig.5. Across these metrics, our method demonstrates a significant advantage, which indicates the high effectiveness of our method in mitigating LVLM's hallucinations.

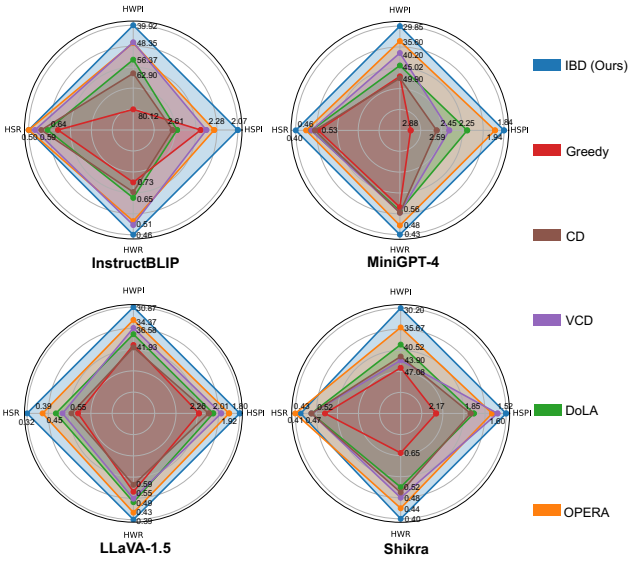


Figure 5. Evaluation results assisted by GPT-4, including 4 metrics: the number of hallucinated sentences per image (HSPI), the number of hallucinated words per image (HWPI), the ratio of hallucinated sentences (HSR), and the ratio of hallucinated words (HWR). Lower values indicate fewer hallucinations.

Table 3. Evaluation results assisted by GPT-4V, including two metrics: correctness (C) and detailedness (D). Higher values indicate fewer hallucinations.

Method	InstructBLIP		MiniGPT-4		LLaVA-1.5		Shikra	
	C	D	C	D	C	D	C	D
Beam Search	5.4	5.2	4.5	4.9	5.9	5.2	5.2	5.0
CD	5.4	5.4	4.8	5.2	6.2	5.2	5.4	5.2
OPERA	6.0	5.6	5.2	5.0	6.3	4.9	5.9	5.0
VCD	6.0	5.8	5.5	5.2	6.4	5.0	6.1	5.1
DoLa	5.8	5.9	5.0	5.0	6.3	5.3	5.8	5.2
IBD	**6.3**	**6.1**	**5.8**	**5.4**	**6.7**	**5.5**	**6.6**	**5.8**

GPT-4V Assisted Evaluation. We further evaluate LVLM's description quality on the MSCOCO dataset by inputting both the image and the generated description into GPT-4V. Concretely, utilizing the method proposed by [14], we instruct GPT-4V to score these descriptions on two metrics, Accuracy and Detailedness, on a scale from 0 to 10.

Table 4. Evaluation of IBD's different components on CHAIR metric. DA and APC refer to dynamic adjustment (Sec.3.3) and adaptive plausibility constraint (Sec.3.4) respectively.

	InstructBLIP		MiniGPT-4		LLaVA-1.5		Shikra	
Method	C_S	C_I	C_S	C_I	C_S	C_I	C_S	C_I
IBD	**15.0**	**6.2**	**21.0**	**7.4**	**12.7**	**4.5**	**13.2**	**5.2**
w/o DA	19.5	7.0	23.3	8.0	18.0	5.5	18.9	5.9
w/o Prompt P	17.2	6.5	22.0	7.8	15.8	5.0	16.4	5.6
w/o APC	15.8	6.4	22.0	7.7	13.5	4.6	14.1	5.5
w/o I_{sim} in Eq.4	15.7	6.4	21.8	7.6	13.6	4.8	13.9	5.5
w/o I_{con} in Eq.4	18.3	6.8	22.5	7.9	16.5	5.1	18.0	5.8
w/ POS-based I_{con}	16.7	6.5	21.8	7.7	14.5	4.8	15.7	5.6

The results are presented in Table.3. In comparison to other methods, our method achieves a significant enhancement. Particularly on the Shikra baseline model, our method surpasses the runner-up by 11.5% in the Detailedness metric. Such gains can be attributed to our method's enhanced exploitation of image information, which enables the generation of descriptions that are more comprehensive with fewer information omissions.

4.3. Ablation Study

Evaluation of Different Components. We conduct an ablation study based on the CHAIR metric to examine the effectiveness of each design component. Specifically, as shown in Table.4, the evaluated components include: (1) dynamic adjustment (DA) for decoding (Sec.3.3), (2) prompt P for finetuning (Sec.3.3); and (3) the adaptive plausibility constraint (APC) as in Eq.7 and Eq.8. Additionally, we evaluate two indicators I_{sim} and I_{con} used within the dynamic adjustment method (Eq.4). Removing any of these components would result in a notable decrease in performance. In Eq.6, we compute I_{con} by using the distances between predictions from different layers, rather than directly employing a POS tagging tool, which may suffer from semantic ambiguities (details in sec.3.3). Our experiments confirm this notion, with the use of a POS-based I_{con} tagged by spacy [12] significantly degrading performance. These results demonstrate the rationality and effectiveness of different design components in our method.

Settings for Hyper-parameter ϵ. As shown in Eq.3, we introduce a factor added to the QK-multiplication result to bias model $\hat{\theta}$ towards image information. This factor is equal to a hyper-parameter ϵ when the key corresponds to an image token. In Fig.6, using the CHAIR metric, we present the performance when using different values of ϵ. The results indicate suboptimal performance when ϵ is either too small or too large. This is attributed to the fact that an excessively small ϵ results in an insufficient bias towards the image, while an overly large ϵ introduces too much noise, thus impairing the model's performance. When $1.5 < \epsilon < 3$, the performance can keep relatively stable.

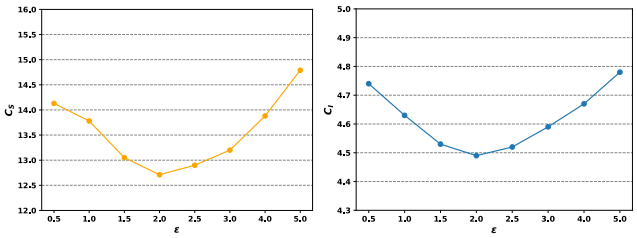


Figure 6. Evaluation results when using different ϵ. The left figure shows the results of the C_S metric in CHAIR, and right figure shows the results of the C_I metric in CHAIR. Smaller values indicate fewer hallucinations.

5. Conclusion

This paper proposes image-biased decoding (IBD), a novel decoding method that aims at alleviating the issue of hallucinations in LVLMs. Our approach involves conducting a prediction contrast between the original model and an image-biased model to amplify the accurate information associated with image content, thereby improving the factuality of the generated text. In addition, we design a dynamic adjustment strategy that flexibly handles different types of vocabulary. Experimental results demonstrate the high effectiveness of IBD in addressing LVLM's hallucination challenges. We hope our innovative method and detailed statistical analysis can provide valuable insights for future research in this field.

References

[1] 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. 2

[2] Assaf Ben-Kish, Moran Yanuka, Morris Alper, Raja Giryes, and Hadar Averbuch-Elor. Mocha: Multi-objective reinforcement mitigating caption hallucinations. *arXiv preprint arXiv:2312.03631*, 2023. 2

[3] Nicolas Boulanger-Lewandowski, Yoshua Bengio, and Pascal Vincent. Audio chord recognition with recurrent neural networks. In *ISMIR*, pages 335–340. Curitiba, 2013. 2, 7

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1, 2

[5] 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. 6

[6] Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. Halc: object hallucination reduction

via adaptive focal-contrast decoding. In *Proceedings of the 41st International Conference on Machine Learning*, pages 7824–7846, 2024. 2

[7] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*, 2023. 2, 4, 5, 7

[8] 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, 2023. 2, 6

[9] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2018. 2

[10] Yudong Han, Liqiang Nie, Jianhua Yin, Jianlong Wu, and Yan Yan. Visual perturbation-aware collaborative learning for overcoming the language prior problem. *arXiv preprint arXiv:2207.11850*, 2022. 1, 3

[11] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2019. 1, 2, 7

[12] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017. 8

[13] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025. 1

[14] 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. *arXiv preprint arXiv:2311.17911*, 2023. 2, 6, 7

[15] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023. 1

[16] Jae Myung Kim, A Koepke, Cordelia Schmid, and Zeynep Akata. Exposing and mitigating spurious correlations for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2584–2594, 2023. 2

[17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 7

[18] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023. 2

[19] 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. *arXiv preprint arXiv:2311.16922*, 2023. 1, 2, 7

[20] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1, 2

[21] 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. 2, 4, 5, 7

[22] 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. 1, 2, 3

[23] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 1(2):9, 2023. 2

[24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 1, 2, 6

[25] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024. 1

[26] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Erix Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. *arXiv preprint arXiv:2311.03356*, 2023. 1

[27] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018. 2

[28] Surat Teerapittayanon, Bradley McDanel, and Hsiang-Tsung Kung. Branchynet: Fast inference via early exiting from deep neural networks. In *2016 23rd international conference on pattern recognition (ICPR)*, pages 2464–2469. IEEE, 2016. 5

[29] 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, 2

[30] Bin Wang, Fan Wu, Xiao Han, Jiahui Peng, Huaping Zhong, Pan Zhang, Xiaoyi Dong, Weijia Li, Wei Li, Jiaqi Wang, et al. Vigc: Visual instruction generation and correction. *arXiv preprint arXiv:2308.12714*, 2023. 2

[31] Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. Mllm can see? dynamic correction decoding for hallucination mitigation. *arXiv preprint arXiv:2410.11779*, 2024. 2

[32] Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint*

arXiv:2308.15126, 2023. 2

[33] Lei Wang, Jiabang He, Shenshen Li, Ning Liu, and Ee-Peng Lim. Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites. *arXiv preprint arXiv:2312.01701*, 2023. 2, 7

[34] Teng Wang, Jinrui Zhang, Junjie Fei, Yixiao Ge, Hao Zheng, Yunlong Tang, Zhe Li, Mingqi Gao, Shanshan Zhao, Ying Shan, et al. Caption anything: Interactive image description with diverse multimodal controls. *arXiv preprint arXiv:2305.02677*, 2023. 1

[35] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 1

[36] Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Vary: Scaling up the vision vocabulary for large vision-language models. *arXiv preprint arXiv:2312.06109*, 2023.

[37] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023. 1

[38] Jiarui Xu, Xingyi Zhou, Shen Yan, Xiuye Gu, Anurag Arnab, Chen Sun, Xiaolong Wang, and Cordelia Schmid. Pixel aligned language models. *arXiv preprint arXiv:2312.09237*, 2023. 2

[39] 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. 2, 7

[40] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity, 2023. 1

[41] 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. *arXiv preprint arXiv:2311.13614*, 2023. 2

[42] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023. 2

[43] Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*, 2023. 2

[44] 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. 1, 2, 4

[45] 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, 6

[46] Lanyun Zhu, Tianrun Chen, Deyi Ji, Jieping Ye, and Jun Liu. Llafs: When large language models meet few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3065–3075, 2024. 1