Bi-DCA: Bi-directional Dual Contrastive Adapting for Alleviating Hallucination in Multimodal Large Language Models

Anonymous ACL submission

Abstract

Multimodal Large Language Models (MLLMs) demonstrate excellent performance across various multimodal tasks. However, they still tend to generate text with hallucinations in certain scenarios. Previous efforts to alleviate hallucinations approach this issue from fine-tuning, dataset, and inference perspectives. Despite these efforts, there are two existing challenges in MLLMs particularly the confusing image objects and generating persistent hallucinations. In this paper, we propose a novel training-free method called **Bi**-directional **D**ual **C**ontrastive **A**dapting (Bi-DCA) to alleviate the hallucinations in MLLMs that can integrate seamlessly into the existing decoding methods. We first design a bi-directional attention mechanism to expand the visual receptive field to address the problem of confusing image objects. Building on this, to alleviate the persistent hallucinations in generated sentences, we propose a dual contrastive adapting strategy to enhance the positive effect of images during the next token prediction stage. We conduct extensive experiments using various evaluation methods and benchmarks for hallucination. The experimental results demonstrate that our Bi-DCA not only alleviates the above challenges but achieves superior performance compared with previous methods.

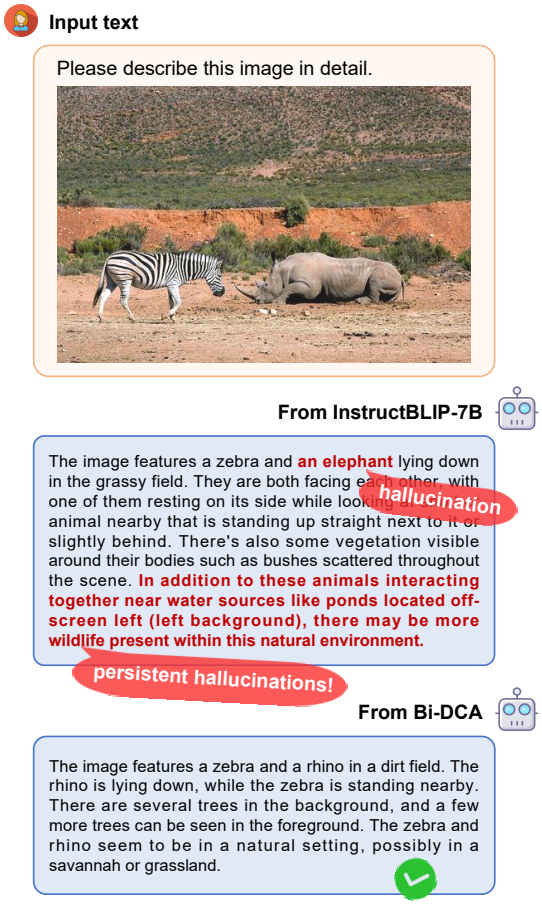

Figure 1: Illustration of the hallucinations when describing the image.

1 Introduction

Multimodal Large Language Models (MLLMs) demonstrate their strong comprehension and generation abilities in many tasks (Cho et al., 2022; Shao et al., 2023; Kim et al., 2023). Despite their impressive performance, MLLMs are found to struggle with the "hallucinations" problem. This means their output responses are often unrelated to the inputs, especially the visual content, leaving significant challenges for practical applications such as medical imaging (Ma et al., 2024) and autonomous driving (Chib and Singh, 2023).

Previous approaches mitigate hallucination by fine-tuning with specifically constructed training data (Ben-Kish et al., 2023; Liu et al., 2023a) or employing reinforcement learning with human feedback (Gunjal et al., 2024), which require external annotation costs and computational resources. Consequently, researchers begin exploring hallucination mitigation methods that do not require additional training. Opera (Huang et al., 2024a) optimizes the inference process by statistically analyzing hallucination patterns from self-attention

maps. At the same time, VCD (Leng et al., 2023) mitigates hallucinations through visual contrastive decoding caused by over-reliance on linguistic priors and statistical biases.

Despite their effectiveness, these methods still face two main challenges: *(i) Confusing image objects.* During the inference stage of MLLMs, the use of causal attention leads to an incomplete receptive field of the image, disrupting the integrity of image features and causing confusion among objects with similar local features. As shown in Figure 1, the response from InstructBLIP misidentifies a rhino as an elephant due to their similar local texture features. *(ii) Generating persistent hallucinations.* As the length of the generated sequence increases, the positive effect of the image on the next token prediction phase gradually diminishes. Thus the model tends to generate a lot of persistent hallucinations in the end. As shown in Figure 1, the response from InstructBLIP introduced by "In addition" in the latter part is significantly inconsistent with the content of the image.

In this paper, we propose a novel approach called **Bi**-directional **D**ual **C**ontrastive **A**dapting (Bi-DCA) that integrates seamlessly into the existing decoding methods to address above challenges and alleviate the hallucinations. Our method does not require additional training or data, it mainly focuses on two innovative mechanisms:

(i) Expanding the Visual Receptive Field. To address the confusing objects, we are inspired by the need to improve the ability to capture directional dependencies within an image. By utilizing directional masks based on relation-aware self-attention, we encode directional information and create a bi-directional attention mechanism when calculating image patches during the inference stage. As our inference stage shown in Figure 2, when the orange-marked patch is computed, it allows the model to consider both forward and backward dependencies, which we call the full visible state. So that it effectively expands the visual receptive field and improves the integration of visual information into the inference process.

(ii) Dual Contrastive Adapting in Predictions. Based on the first step, to alleviate the persistent hallucinations in generated sentences, we propose a dual contrastive adapting strategy to enhance the positive effect of the image on the next token prediction phase. By incorporating multiple prediction scores derived from different visual states such as partial visible, full visible, and full visible rotated

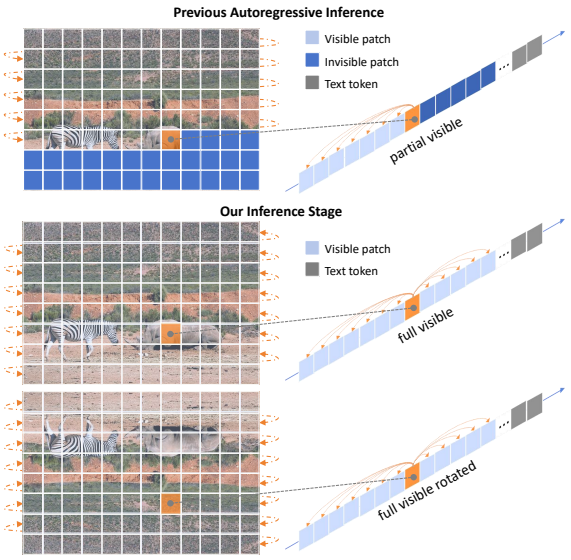

Figure 2: Illustration of expanding the visual receptive field and the different visual state features.

state which are shown in Figure 2, we can adapt the distribution of output score when predicting each token. Specifically, we take the partial visible state as the foundation and combine it with the visual receptive field information provided by the full visible state and the visual spatial information provided by the full visible rotated state, and then, involve them in the above prediction process. Hence, dual contrastive adapting in predictions acts as a mechanism for supplementing visual features and alleviating persistent hallucinations.

During the experiments on various MLLMs and decoding methods, we evaluate the performance of Bi-DCA in alleviating hallucination tasks using various evaluation methods and benchmarks including CHAIR(Rohrbach et al., 2018), POPE(Li et al., 2023b), MME(Yin et al., 2023a), and GPT-4(Achiam et al., 2023). The results indicate that our method has significantly improved performance compared to previous approaches.

In summary, our contributions are as follows:

- We design a bi-directional attention mechanism to address the confusing image objects by expanding the visual receptive field.
- We propose a dual contrastive adapting in predictions strategy for supplementing visual features and alleviating persistent hallucinations.
- Through comprehensive experiments on various benchmarks, we demonstrate the effectiveness of our proposed training-free Bi-DCA.

2 Related Work

2.1 Multimodal Large Language Models

Multimodal Large Language Models (MLLMs), also referred to as Large Vision Language Models (LVLMs), aim to enhance the visual capabilities of Large Language Models (LLMs). The integration of visual and textual modalities is mainly based on off-the-shelf pre-trained unimodal models (Bai et al., 2024). Specifically, these MLLMs usually incorporate a learnable interface between pre-trained visual encoders and LLMs, which can be further categorized into *(i) projection layer-based interface* and *(ii) learnable query-based interface.* Projection layer-based methods, which are widely implemented in models such as LLaVA (Liu et al., 2023b) and Shikra (Chen et al., 2023), involve training a linear projection layer or a Multi-Layer Perceptron (MLP) module to transform extracted visual features. On the other hand, learnable query-based methods, exemplified by Q-Former (Li et al., 2023a), as utilized in InstructBLIP (Dai et al., 2024) and MiniGPT-4 (Zhu et al., 2023), employ a set of learnable query tokens to capture visual signals through cross-attention mechanisms. Both types of interfaces aim to map pre-trained visual features into the input space of pre-trained LLMs, thereby facilitating the integration of visual and textual information.

In our paper, we conduct experiments on the four aforementioned MLLMs to validate the robustness of our proposed Bi-DCA.

2.2 Hallucination in MLLMs

The hallucination of MLLMs generally refers to the problem where the generated text response is not consistent with the given visual content (Huang et al., 2024b). State-of-the-art studies in this field primarily focus on object hallucination, which can be categorized into object-level category and attribute-level category. The object-level meanings identify nonexistent object categories or incorrect categories in the given image, and attribute-level refers to the descriptions of the attributes on these objects such as color, position, etc. are wrong. Current methods for evaluating hallucinations in MLLMs focus on assessing the cognitive performance of the model, with two primary aspects: non-hallucinatory generation and hallucination discrimination. The former involves a detailed analysis of the hallucinatory elements in the text response and quantifying their proportion. The latter requires a binary judgment of whether the response comprises any hallucinatory content.

In our paper, we alleviate both two object hallucinations in the generated text and comprehensively discuss these evaluating approaches based on our Bi-DCA.

2.3 Decoding Method in Language Models

In constructing language models, the decoding method plays a crucial role in the text generation process. These methods are essential for ensuring the accuracy, relevance, and fluency of the generated text. A basic decoding method is the greedy search, which selects the word with the highest probability at each step. Although this method is computationally efficient, it often results in monotonous and less diverse content. In contrast, beam search (Graves, 2012; Lee et al., 2009) maintains a certain number of candidate sequences at each step and selects the optimal sequence from them, thereby improving the quality and diversity of the generated text. Nucleus sampling (Holtzman et al., 2019) involves randomly selecting from a set of words, it is not simply choosing a fixed number of words with the highest probabilities, but determining the number of words to select based on a cumulative probability value p. It achieves an effective balance between randomness and text relevance in text generation by adjusting the number of selected words.

In our paper, the proposed Bi-DCA can integrate seamlessly into the above decoding methods, thus, it can be represented as greedy-based Bi-DCA, beam-based Bi-DCA, and sample-based Bi-DCA, respectively.

3 Method

Our core objective is to alleviate hallucinations by expanding the receptive field for visual features while achieving dual contrastive adapting in the prediction phase. Our method can be seamlessly integrated into existing mature decoding approaches. In this section, we introduce the Inference Process of MLLMs, followed by a detailed explanation of Expanding the Visual Receptive Field and Dual Contrastive Adapting in Predictions.

3.1 Inference Process of MLLMs

The key to the inference process is the visual encoder and the decoder of the large language model. Specifically, the model receives a given

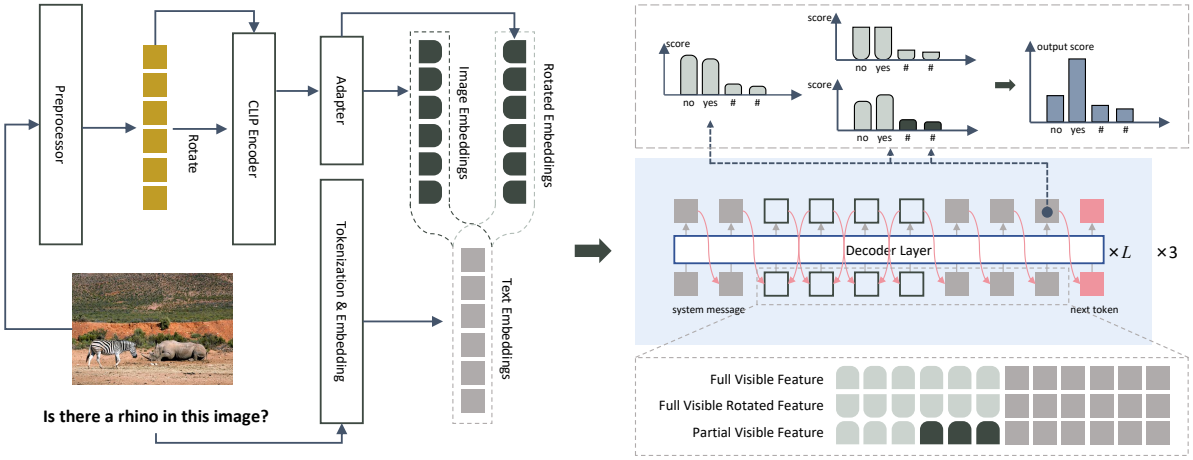

Figure 3: Illustration of our method. Given an image and text input, the model first extends the receptive field of the image and combines image and text features with attention in two directions to form distinct features. Subsequently, a dual contrastive decoding mechanism is employed to control the output scores for predicting the next token.

image V and text T as inputs, the text is transformed into a fixed dimensional vector representation $X_T = \{x_N, x_{N+1}, x_{N+M-1}\}$ after an embedding layer, the visual encoder encodes the image as $X_V = \{x_0, x_1, x_{N-1}\}$ which fuses with the text vectors through an alignment such as a linear layer or a Qformer, and the fused vector $X = X_V \oplus X_T = \{x_0, x_1, ..., x_{N+M-1}\}$ is used as a prediction for the start of the target sequence. Here N and M are the length of visual and textual tokens that are a fixed value in most cases.

Then, the decoder enters the loop generation phase, where for each time step t, it generates a new word vector from a predicted score logit$(\cdot)$ and aligns it to the next position of the target sequence. This process of auto-regressive can be formulated as:

$$p(x_t|x_{<t}) = \text{softmax}(\text{logit}(x_t|x_{<t})), \quad (1)$$

where x_t is the t-th token which is conditioned on all previous tokens $x_{<t}$.

After getting the probability distribution of the next token, several decoding strategies are usually utilized to obtain the final output, such as greedy search, beam search, sampling, *etc*. Our method can be efficiently and easily added to these decoding methods.

3.2 Expanding the Visual Receptive Field

Inspired by Shen et al. (2018), to enhance the information integrity of the image in the decoding stage, we utilize different positional masks to encode the directional information in it based on Relation-aware Self-Attention (Shen et al., 2018). It can be represented by the following equation:

$$z_i = \sum_{j=1}^{n} \alpha_{ij}(x_j W^V + a_{ij}^V), \quad (2)$$

$$\alpha_{ij} = \text{softmax}(e_{ij}), \quad (3)$$

$$e_{ij} = \frac{x_i W^Q (x_j W^K + a_{ij}^K)^T}{\sqrt{d_z}}, \quad (4)$$

where $a_{ij}^V, a_{ij}^K \in \mathbb{R}^{d_a}$ are the edge between input elements x_i and x_j, and these representations can be shared across attention heads and $d_a = d_z$.

We modify Eq. (4) to propagate directional information to the sublayer output:

$$e_{ij} = \frac{x_i W^Q (x_j W^K + a_{ij}^K)^T}{\sqrt{d_z}} + \mathcal{M}_{ij}, \quad (5)$$

where $\mathcal{M} \in \{0, -\infty\}^{n \times n}$, n is the image patch numbers. In this paper, we use two positional masks, i.e., forward mask $\mathcal{M}^{fw}$ and backward mask $\mathcal{M}^{bw}$ when combining different modality features. Together they form a bi-directional mask through concatenation.

$$\mathcal{M}_{ij}^{fw} = \begin{cases} 0 & \text{if } i < j, \\ -\infty & \text{otherwise}, \end{cases} \quad (6)$$

$$\mathcal{M}_{ij}^{bw} = \begin{cases} 0 & \text{if } i > j, \\ -\infty & \text{otherwise}. \end{cases} \quad (7)$$

In forward mask $\mathcal{M}_{ij}^{fw}$, there is the only attention of later token j to early token i, and vice versa

in the backward mask. As shown in Figure 3, we abstractly show the schematic of the bi-directional mask added to the input of the decoding layer, especially in the image patch region. The idea of using bi-directional attention is inspired by different types of image and text feature processing. Causal attention loses the information of the following patches when processing images. Unlike textual features which only need to focus on the tokens before the current token, it needs to encode long-range dependency from different directions, so that expanding the visual receptive field.

3.3 Dual Contrastive Adapting in Predictions

We have discussed the methods to enhance the visual perceptual field. We further propose novel approaches to compensate for the spatial information and combine both to form a dual contrastive adapting strategy to achieve dynamic adaptation in the prediction phase.

The original MLLMs assign the highest probability score to the wrong token when it outputs factually incorrect information, in which case we observe that the score of the correct token is close to the highest probability score. It suggests that the model is less confident in the current decision. Whereas the output score has a large difference between the token with the highest probability score and the token with the second highest probability score when it outputs the correct outcome. Since the MLLMs are trained with causal attention using only forward masks, and when using backward masks, it does not learn the spatial information in the opposite direction. So we use the rotated image to compensate for this part of the spatial information.

Based on the above analysis, our goal is to go against the decision scores in the incorrect case, both in terms of visual perceptual field and spatial information completeness.

To operationalize this objective, in addition to the original state score $p(x_t|x_{<t})$, which we call the partial visible score calculated by original image and forward mask $\mathcal{M}^{fw}$, denoted as $e(x_t, \phi_\rho, \phi_\theta)$, we introduce two other prediction scores, namely full visible score $e(x_t, c_\rho, \phi_\theta)$ and full visible rotated score $e(x_t, c_\rho, c_\theta)$. where c_ρ represents the Expansion of Visual Receptive Field introduced in Sec. 3.2, and c_θ stands for spatial semantic information obtained after rotational correction of the image. In this paper we set θ to 180, meaning that the original image is rotated 180 degrees.

As shown in Figure 3, these scores are derived from the combination of two different image features and positional masks, which are then processed through the decoder layer. The final next token score for step t can be derived from the following equation:

$$\begin{aligned} e_t = \alpha_1 e(x_t, \phi_\rho, \phi_\theta) \\ + \alpha_2(e(x_t, c_\rho, \phi_\theta) - e(x_t, \phi_\rho, \phi_\theta)) \\ + \alpha_3(e(x_t, c_\rho, c_\theta) - e(x_t, c_\rho, \phi_\theta)), \end{aligned} \quad (8)$$

where $\alpha_1, \alpha_2, \alpha_3$ are hyper-parameters in $[0, 1]$. we set $\alpha_1 = 1$, $\alpha_2 = 0.5$, $\alpha_3 = 0.5$ throughout the paper. Larger α_2 entails more visual receptive field and larger α_3 means more visual spatial supplementary information. However, their values are not necessarily better when larger, as this can offset the original state scores and cause the model to favor additional states.

To tackle the aforementioned issue, following Li et al. (2022); Leng et al. (2023), we utilize an adaptive plausibility constraint $\mathcal{V}_{head}$ that exploits the confidence level of the original state score to restrict the effect of the rest objective:

$$\begin{aligned} \mathcal{V}_{head} = \{x_t \in \mathcal{V} : \\ e(x_t, c_\rho, \phi_\theta) \geq \max e(x_t, c_\rho, \phi_\theta) + \log(\beta)\}, \\ e_t = 0, \text{if } x_t \notin \mathcal{V}_{head}, \end{aligned} \quad (9)$$

where β is a hyper-parameter ranging from 0 to 1 that controls the truncation of the next token score distribution. Larger β signifies more aggressive truncation, retaining only the tokens with the highest probabilities.

By incorporating adaptive plausibility constraint into Eq. (8), we retained tokens with higher predicted probabilities, thereby altering the distribution of the final output scores. This increases the confidence of the model in its output decisions and reduces the emergence of low-probability scores. Then we apply existing mature decoding approaches such as search-based greedy search, beam search, and sample-based nucleus sampling to optimize e_t and select a token with a higher probability.

4 Experiments

This section provides a detailed overview of our experimental validation of decoding strategies employed in different MLLMs.

		LLaVA-1.5			Shikra			InstructBLIP			MiniGPT-4		
		CHAIRs↓	CHAIRi↓	Len	CHAIRs↓	CHAIRi↓	Len	CHAIRs↓	CHAIRi↓	Len	CHAIRs↓	CHAIRi↓	Len
Opera		45.6	13.3	94.2	52.2	13.8	99.5	48.8	14.9	91.6	26.8	8.9	63.0
VCD		48.8	14.1	98.0	56.2	15.1	101.2	46.4	14.7	96.8	33.2	10.5	83.3
Greedy	✗	46.4	12.4	97.8	54.8	14.8	101.5	49.3	22.6	108.8	**32.4**	**10.1**	83.5
	✓	**43.6**	**11.9**	94.2	**53.6**	**13.5**	99.4	**48.4**	**14.6**	93.6	32.6	10.9	91.8
Beam	✗	49.4	14.0	96.9	52.2	13.8	99.5	56.8	15.3	98.2	**31.2**	**10.0**	78.5
	✓	**42.2**	**11.7**	93.3	**50.4**	**13.1**	93.3	**44.4**	**13.8**	97.8	32.4	10.7	81.6
Sample	✗	54.2	15.8	99.8	60.2	16.4	102.3	50.0	24.5	118.4	33.8	10.6	83.9
	✓	**43.2**	**11.5**	94.4	**56.2**	**15.1**	101.2	**45.4**	**14.4**	96.1	**29.6**	**9.0**	88.4

Table 1: CHAIR metrics across four different MLLMs. ✗symbol represents the original decoding method, while ✓indicates our proposed Bi-DCA based on the respective decoding strategy. Len represents the average length of the generated sentences and is provided for reference. The best performances within each setting are bolded.

4.1 Settings

4.1.1 Baselines & Dataset

Following Huang et al. (2024a), we evaluate the effectiveness of Bi-DCA on four MLLMs, including LLaVA-1.5-7B (Liu et al., 2023b), Shikra-7B (Chen et al., 2023), InstructBLIP-7B (Dai et al., 2024), and MiniGPT-4-7B (Zhu et al., 2023). All the models employ pre-trained LLMs, efficient image encoders, and different visual feature alignment modules. The first is LLaVA-1.5 and Shikra which use a linear MLP as the image-text feature alignment module, the numbers of the image patches are 576 and 256. While the InstructBLIP and Minigpt4 both map the image features into the textual space using the Q-former (Li et al., 2023a) structure. The pre-trained LLMs they used are LLaMA-7B (Touvron et al., 2023) and Vicuna-7B (Chiang et al., 2023) respectively, and the image encoders used are CLIP ViT (Radford et al., 2021) or EVA-CLIP ViT (Fang et al., 2023). During the inference phase, we select five decoding methods as baseline approaches. These include three common strategies: greedy search, beam search, and nucleus sampling, as well as two methods designed to alleviate hallucinations: OPERA (Huang et al., 2024a) and VCD (Leng et al., 2023).

We conduct experiments on the MSCOCO dataset, in which the images contain 80 categories and corresponding annotations. Specifically, following Huang et al. (2024a), we select 500 images from the COCO14 (Lin et al., 2014) validation set and then prompt different models to obtain descriptions of the input images and evaluate the performance of the models by assessing the quality of the outputs.

4.1.2 Implementation Details

we utilize the default settings for these models and decoding methods during the experiments. Specifically, we set the beam search parameter N_{beam} to 5 and the $top\text{-}p = 0.9$ for nucleus sampling. For the VCD, we set $\alpha = 1$, $\beta = 0.1$, $\gamma = 0.1$. And for the OPERA, we configure them as follows: $N_{beam} = 5$, $\theta = 50$, $N_{can} = 5$, $\alpha = 1$, $\beta = 5$ and $r = 15$. Unless otherwise specified, we set $\alpha 1 = 1$, $\alpha 2 = 0.5$, $\alpha 3 = 0.5$, $\theta = 180$, and $\beta = 0.5$ across all models in our Bi-DCA.

4.2 Experimental Results

we evaluate the performance of Bi-DCA in alleviating hallucination tasks using various evaluation methods, including two primary hallucination evaluation approaches: *(i)* Assessing the ability of non-hallucinatory content generation. *(ii)* Evaluating the ability of hallucination discrimination.

4.2.1 Results on CHAIR

Evaluating non-hallucinatory generation is to measure the proportion of hallucinated content in the outputs. CHAIR (Rohrbach et al., 2018) targets evaluating object hallucinations of models in describing images by quantifying differences of objects between model generation and ground truth. It comprises two metrics dimensions: CHAIRs calculated at the sentence level, and CHAIRi calculated at the object level. These variables can be expressed using the following formulas:

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

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

Model	Setting	Decoding	Accuracy↑	Precision↑	Recall↑	F1 Score↑
LLaVA-1.5	Random	✗$_{Avg}$	0.875	**0.899**	0.858	0.877
		✓$_{Avg}$	**0.884**	0.875	**0.904**	**0.890**
	Popular	✗$_{Avg}$	0.848	0.845	0.858	0.850
		✓$_{Avg}$	**0.850**	0.816	**0.904**	**0.858**
	Adversarial	✗$_{Avg}$	0.767	0.710	0.904	0.795
		✓$_{Avg}$	**0.779**	**0.733**	0.879	**0.799**
Shikra	Random	✗$_{Avg}$	0.821	**0.949**	0.691	0.800
		✓$_{Avg}$	**0.829**	0.944	**0.712**	**0.811**
	Popular	✗$_{Avg}$	0.809	0.904	0.692	0.784
		✓$_{Avg}$	**0.819**	**0.906**	**0.711**	**0.796**
	Adversarial	✗$_{Avg}$	0.792	**0.867**	0.690	0.769
		✓$_{Avg}$	**0.799**	0.863	**0.711**	**0.780**
InstructBLIP	Random	✗$_{Avg}$	0.877	**0.916**	0.841	0.876
		✓$_{Avg}$	**0.904**	0.910	**0.904**	**0.906**
	Popular	✗$_{Avg}$	0.816	**0.804**	0.842	0.821
		✓$_{Avg}$	**0.817**	0.771	**0.904**	**0.832**
	Adversarial	✗$_{Avg}$	**0.788**	**0.767**	0.835	0.799
		✓$_{Avg}$	0.782	0.728	**0.902**	**0.805**
MiniGPT-4	Random	✗$_{Avg}$	0.743	0.814	0.660	0.728
		✓$_{Avg}$	**0.794**	**0.909**	**0.668**	**0.769**
	Popular	✗$_{Avg}$	0.688	0.702	0.666	0.683
		✓$_{Avg}$	**0.741**	**0.784**	**0.667**	**0.720**
	Adversarial	✗$_{Avg}$	0.669	0.672	0.669	0.669
		✓$_{Avg}$	**0.721**	**0.742**	**0.680**	**0.708**

Table 2: POPE metrics across four different MLLMs. Due to space constraints, we use ✗$_{Avg}$ to denote the average results of beam search, greedy search, and nucleus sampling in different settings. ✓$_{Avg}$ reflects the average of our method based on these three methods, with the best result for each setting highlighted in bold.

In our experiments on the MSCOCO dataset, specifically aimed at obtaining detailed descriptions of input images, we utilized the same prompt "*Please describe this image in detail.*" to get responses from different MLLMs.

As shown in Table 1, our CHAIR results on different MLLMs and baseline methods demonstrate a noticeable observation: our proposed Bi-DCA exhibits superior robustness. Specifically, our method outperforms the baselines across four different models. For the MiniGPT-4 model, our method based on nucleus sampling achieves the best results, with the outcomes of other decoding strategies also comparable. In models other than MiniGPT-4, the performance exceeds the baselines by 5% or more. This indicates that it plays a crucial role in simultaneously enhancing the visual receptive field and improving the robustness against perturbations, thereby reducing the occurrence of object hallucinations.

4.2.2 Results on POPE

The Hallucination discrimination evaluation approach aims to assess the hallucination discrimination ability of MLLMs. The methods that follow this approach typically adopt a question-answering format, posing inquiries to MLLMs consisting of descriptions that agree or conflict with the provided content (Bai et al., 2024). POPE (Li et al., 2023b) designs binary (Yes-or-No) questions about object presence in images such as *"Is there a <object> in the image?"* to evaluate the hallucination discrimination ability of MLLMs. The objects asked in questions are selected under three distinct sampling strategies: random (selecting random absent objects), popular (choosing the most frequent objects in the dataset but absent in the current image), and adversarial (selecting absent objects often co-occurring with present ones). As shown in Table 2, our proposed Bi-DCA demonstrates a robust enhancement in the performance of four MLLMs across various settings. The consistent improvements in accuracy, recall, and F1 scores, especially under challenging settings like Adversarial and Popular, underline the effectiveness of the hallucination discrimination ability of Bi-DCA. The overall performance improvement of minigpt4 is attributed to the comprehensive improvement in accuracy, precision, and recall, while the performance of the other three models is mainly driven by accuracy and recall. Notably, InstructBLIP and MiniGPT4 exhibit more significant enhancements in their F1 metrics compared to LLaVA1.5 and SHIKRA.

4.2.3 Results on MME

We select four subsets related to hallucination from MME benchmark (Yin et al., 2023a) for experiments, specifically existence, count, position, and color. These subsets surpass the evaluation scope of POPE, providing a more comprehensive understanding of our proposed Bi-DCA. The results are shown in Table 3. Overall, it indicates that our method contributes to a consistent enhancement of model performance when alleviating hallucination at the object and attribute levels. In addition, Bi-DCA shows a significant improvement in its ability to discern and alleviate hallucinations at the attribute level, particularly regarding positional hallucinations. This precisely demonstrates that the integration of spatial information in our method has achieved the desired effect. However, the scores for position metrics are generally lower than the other three metrics, indicating that the reasoning capabilities of MLLMs regarding position still need improvement. When comparing different decoding methods, our Bi-DCA shows the most significant score improvement with nucleus sampling, achieving an average increase of 75.55% on LLaVA-1.5, SHIKRA, and MiniGPT4. In contrast, the improve-

Model	Decoding	Object-level		Attribute-level		Total Scores↑
		Existence↑	Count↑	Position↑	Color↑	
LLaVA-1.5	sample	175.00	110.00	95.00	135.00	515.00
	Ours	**190.00**	**130.00**	**123.33**	**165.00**	**608.33**
	greedy	195.00	146.67	121.67	170.00	633.33
	Ours	**195.00**	**153.33**	**131.67**	**170.00**	**650.00**
	beam	195.00	118.33	110.00	150.00	573.33
	Ours	**195.00**	**128.33**	**110.00**	**150.00**	**583.33**
Shikra	sample	165.00	51.67	45.00	103.33	365.00
	Ours	**175.00**	**80.00**	**61.67**	**123.33**	**440.00**
	greedy	**195.00**	61.67	53.33	93.33	403.33
	Ours	175.00	**70.00**	**63.33**	**115.00**	**423.33**
	beam	**195.00**	**83.33**	60.00	88.33	**426.67**
	Ours	175.00	78.33	**63.33**	**108.33**	425.00
InstructBLIP	sample	180.00	70.00	**61.67**	110.00	421.67
	Ours	**185.00**	**75.00**	56.67	**115.00**	**431.67**
	greedy	185.00	60.00	50.00	120.00	415.00
	Ours	**185.00**	**65.00**	**53.33**	**125.00**	**428.33**
	beam	185.00	55.00	50.00	120.00	410.00
	Ours	**185.00**	**65.00**	**53.33**	**125.00**	**428.33**
MiniGPT-4	sample	65.00	48.33	25.00	46.67	185.00
	Ours	**95.00**	**61.67**	**66.67**	**95.00**	**318.33**
	greedy	115.00	**56.67**	60.00	85.00	316.67
	Ours	**120.00**	51.67	**71.67**	**93.33**	**336.67**
	beam	95.00	**91.67**	53.33	83.33	323.33
	Ours	**110.00**	71.67	**80.00**	**95.00**	**356.67**

Table 3: MME metrics across four different MLLMs. The best result for each setting is highlighted in bold.

Model		Grammer↑	Fluency↑	Nature↑	PPL_1 ↓	PPL_2 ↓
InstructBLIP	✗	7.34	7.35	6.58	68.08	51.78
	✓	**8.38**	**8.32**	**7.56**	**12.10**	**9.61**
MiniGPT-4	✗	7.77	7.71	7.68	12.51	**9.93**
	✓	**8.10**	**8.07**	**8.01**	**13.00**	10.21
LLaVA-1.5	✗	7.73	7.73	7.63	14.13	11.37
	✓	**8.39**	**8.40**	**8.26**	**13.59**	**10.91**
Shikra	✗	7.75	7.74	7.67	16.61	13.38
	✓	**8.43**	**8.43**	**8.31**	**15.20**	**12.19**

Table 4: GPT-4 assisted evaluation results on COCO14. ✗denotes the average results of beam search, greedy search, and nucleus sampling in different MLLMs. ✓reflects the average results of ours.

ments with other decoding methods are relatively modest.

4.2.4 GPT-4 Assisted Evaluation

Following Zhao et al. (2023); Huang et al. (2024a), to evaluate the quality of the generated text from the traditional NLP perspective, we use GPT-4 to score the image descriptions, specifically assessing their grammar, fluency, and nature from 0-10. Additionally, we adopt perplexity (ppl) to evaluate the generated sentences, with ppl_1 and ppl_2 calculated by gpt-2 and gpt-2-medium models, respectively. We calculate the above metrics based on the CHAIR metric, and Table 4 lists the average scores for various decoding methods and our method. Detailed prompt templates are provided in the Appendix A. The result in Table 4 indicates that the quality of the generated text also improves in various aspects.

Decoding	InstructBLIP		MiniGPT-4		LLaVA-1.5		Shikra	
	C	D	C	D	C	D	C	D
Sample	2.286	3.242	3.816	4.560	3.796	4.584	3.904	4.482
Greedy	3.092	3.502	4.476	4.548	5.092	4.830	4.446	4.526
Beam	4.536	4.900	4.362	4.900	4.462	**5.034**	4.552	**4.970**
Ours	**5.38**	**4.936**	**4.964**	**5.04**	**5.418**	4.988	**5.484**	4.843

Table 5: GPT-4V assisted evaluation results on COCO14. C stands for correctness and D refers to detailness, ours reflects the sampling-based Bi-DCA.

4.2.5 GPT-4V Assisted Evaluation

Following Yin et al. (2023b), we adapt the state-of-the-art gpt-4-vision-preview further to evaluate the presence of hallucinations in the output text. It can compensate for the attribute-level hallucinations that the CHAIR metric cannot detect and has strong capabilities in handling both image and text information. Implementation details and prompt templates are provided in the Appendix B.

The results of four MLLMs using different decoding methods and our methods are presented in Table 5. Overall, our method achieves an 18.6% quality improvement in terms of correctness and is comparable to other leading decoding methods in describing image content in detail. Due to the strong perception and reasoning abilities of gpt-4-vision-preview, which is close to those of humans, the evaluation results to some extent reflect human perspectives on hallucination mitigation.

5 Conclusion

In this paper, we propose a novel training-free method called Bi-directional Dual Contrastive Adapting (Bi-DCA) to alleviate the hallucinations in MLLMs that can integrate seamlessly into the existing decoding methods. It mainly focuses on two innovative mechanisms: Expanding the Visual Receptive Field and Dual Contrastive Adapting in Predictions. First, we employ directional masks to capture the bi-directional dependency of visual information when calculating image patches during the inference stage, which effectively expands the visual receptive field. Building on this, we design a dual contrastive adapting strategy to enhance the confidence of MLLMs in the next token prediction phase, which acts as a mechanism for supplementing visual features and alleviating persistent hallucinations. We conduct comprehensive experiments on various metrics and benchmarks and experimental results show our significant superiority in generating high-quality text and alleviating hallucinations.

Limitations

Our Bi-DCA does not require training and is constrained by the inherent performance of the MLLMs, including the components of the LLM, the visual encoder, and the adapter. When MLLMs respond to questions, the score gap between hallucinated and correct text is small due to their inherent limitations. Although our method can alter this difference and mitigate the hallucinations, the extent of improvement is relatively limited. We hope our approach will inspire researchers so that can prompt further enhancements in model performance. Lastly, due to limited resources, we have not evaluated the most recent larger MLLMs.

Ethics Statement

We affirm that our work here does not exacerbate the biases already inherent in the large language models and does not have ethics problems.

References

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

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

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

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

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

Pranav Singh Chib and Pravendra Singh. 2023. Recent advancements in end-to-end autonomous driving using deep learning: A survey. *IEEE Transactions on Intelligent Vehicles*.

Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, and Mohit Bansal. 2022. Fine-grained image captioning with clip reward. *arXiv preprint arXiv:2205.13115*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18135–18143.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024a. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *Preprint*, arXiv:2311.17911.

Wen Huang, Hongbin Liu, Minxin Guo, and Neil Zhenqiang Gong. 2024b. Visual hallucinations of multimodal large language models. *arXiv preprint arXiv:2402.14683*.

Jae Myung Kim, A Koepke, Cordelia Schmid, and Zeynep Akata. 2023. 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.

Daniel D Lee, P Pham, Y Largman, and A Ng. 2009. Advances in neural information processing systems 22. *Tech Rep*.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2023. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *arXiv preprint arXiv:2311.16922*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. 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.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2022. Contrastive decoding:

Open-ended text generation as optimization. *arXiv preprint arXiv:2210.15097*.

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023b. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. 2024. Segment anything in medical images. *Nature Communications*, 15(1):654.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

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

Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. 2023. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14974–14983.

Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Shirui Pan, and Chengqi Zhang. 2018. Disan: Directional self-attention network for rnn/cnn-free language understanding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

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

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023a. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2023b. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*.

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

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

A Details for GPT-4

GPT-4 Prompt Template
You are an AI language assessment expert tasked with evaluating the quality of text summaries generated by four assistants based on the following criteria: You are required to score the performance of the quality of these four text summaries, Please rate the responses of the assistants on a scale of 0 to 10, where a higher score indicates better performance, according to the following criteria: 1. Grammar: Evaluates whether the text adheres to standard grammatical conventions. 2. Fluency: Assesses the smoothness and coherence of the text. 3. Naturalness: Evaluates how naturally the text reads. Please provide scores for each criterion for each summary containing only four values indicating the scores for Assistant 1 and Assistant 2 respectively. The four scores are separated by a space. Avoid any potential bias and ensure that the order in which the responses were presented does not affect your judgment. **[Assistant 1]** {} **[End of Assistant 1]** **[Assistant 2]** {} **[End of Assistant 2]** Scoring format: Grammar: <Scores of the four answers> Fluency: <Scores of the four answers> Naturalness: <Scores of the four answers>

Table 6: The prompt template for GPT-4.

B Details for GPT-4V

GPT-4V(ision) Prompt Template
You are required to score the performance of two AI assistants in describing a given image. You should pay extra attention to the hallucination, which refers to the part of descriptions that are inconsistent with the image content, such as claiming the existence of something not present in the image or describing incorrectly in terms of the counts, positions, or colors of objects in the image. Please rate the responses of the assistants on a scale of 1 to 10, where a higher score indicates better performance, according to the following criteria: 1: Accuracy: whether the response is accurate with respect to the image content. Responses with fewer hallucinations should be given higher scores. 2: Detailedness: whether the response is rich in necessary details. Note that hallucinated descriptions should not count as necessary details. Please output the scores for each criterion, containing only two values indicating the scores for Assistant 1 and 2, respectively. The two scores are separated by a space. Following the scores, please provide an explanation of your evaluation, avoiding any potential bias and ensuring that the order in which the responses were presented does not affect your judgment. **[Assistant 1]** {} **[End of Assistant 1]** **[Assistant 2]** {} **[End of Assistant 2]** Output format: Accuracy: <Scores of the two answers> Reason: Detailedness: <Scores of the two answers> Reason:

Table 7: The prompt template for GPT-4V(ision).

Specifically, we use 500 images randomly selected from COCO14 and their descriptions generated by various MLLMs and our sample-based Bi-DCA. The prompt provided to the MLLMs is "*Please describe this image in detail.*" To ensure a fair comparison, we follow Yin et al. (2023b); Huang et al. (2024a) and provide gpt-4-vision-preview with both an image and corresponding outputs from different MLLMs, then prompt it to evaluate these generation texts. The template is shown in Table 7 It is asked to score these texts from 0 to 10, based on our defined criteria of correctness and detailness. Correctness refers to the consistency between the text content and the image, while detailness refers to the comprehensiveness of the text description, i.e., whether the image content is completely and accurately described. The score is low if gpt-4-vision-preview determines that the given text does not match the provided image, indicating a hallucination.

C Ablation Study

	α_1	α_2	α_3	β	Accuracy↑	Precision↑	Recall↑	F1 Score↑
	1	0	0	0	83.35	82.70	85.57	84.11
	1	0.5	0.5	0.1	84.36	82.88	87.76	85.25
	1	0.5	0.5	0.3	86.11	84.91	88.82	86.82
β	**1**	**0.5**	**0.5**	**0.5**	**87.90**	**87.00**	**89.95**	**88.45**
	1	0.5	0.5	0.7	88.56	87.59	90.63	89.09
	1	0.5	0.5	0.9	88.64	87.50	90.94	89.19
	0	0.5	0.5	0.5	86.38	84.79	89.65	87.15
	0.1	0.5	0.5	0.5	86.30	84.71	89.58	87.18
	0.3	0.5	0.5	0.5	86.50	85.02	89.58	87.24
α_1	0.5	0.5	0.5	0.5	86.65	85.21	89.65	87.38
	0.7	0.5	0.5	0.5	87.28	86.04	89.88	87.92
	0.9	0.5	0.5	0.5	87.35	86.22	89.80	87.98
	1	**0.5**	**0.5**	**0.5**	**87.90**	**87.00**	**89.95**	**88.45**
	1	0	0.5	0.5	87.51	86.42	89.88	88.12
	1	0.1	0.5	0.5	87.63	86.67	89.80	88.20
	1	0.3	0.5	0.5	87.78	86.86	89.88	88.34
α_2	**1**	**0.5**	**0.5**	**0.5**	**87.90**	**87.00**	**89.95**	**88.45**
	1	0.7	0.5	0.5	87.39	86.18	89.95	88.45
	1	0.9	0.5	0.5	87.16	86.01	89.65	87.80
	1	0.5	0	0.5	87.47	86.25	90.03	88.10
	1	0.5	0.1	0.5	87.51	86.26	90.11	88.14
	1	0.5	0.3	0.5	87.70	86.57	90.11	88.30
α_3	**1**	**0.5**	**0.5**	**0.5**	**87.90**	**89.00**	**89.95**	**88.45**
	1	0.5	0.7	0.5	87.59	86.55	89.88	88.18
	1	0.5	0.9	0.5	87.43	86.45	89.65	88.02

Table 8: Ablation study on POPE Random setting using sample-based Bi-DCA on LLaVA-1.5.

In this section, we present a detailed ablation study of the hyper-parameters, which are introduced in detail in Sec. 3.3. These hyper-parameters include the weight for partial visible scores α_1, the weight for full visible scores α_2, the weight for full visible rotated scores α_3, and the truncation parameter β that controls the distribution of token scores.

Despite minor differences in the optimal hyperparameter settings across various MLLMs, the trends remain consistent. Thus, we conduct our experiment on LLaVA-1.5 using sample-based Bi-DCA. As shown in Tab. 8, for ease of analysis, the first row presents the results under the original conditions. The experimental results demonstrate that our Bi-DCA generally outperforms the baselines. To minimize discrepancies in performance across different MLLMs, we set the default parameters in our paper to $\alpha_1 = 1$, $\alpha_2 = 0.5$, and $\alpha_3 = 0.5$.

Specifically, as β increases, fewer low-score tokens are included in the truncated probability distribution, leading to the output of higher confidence tokens. As shown in Tab. 8, this principle is reflected in the metrics, our F1 score increases when β changes from 0.1 to 0.9. However, to ensure there are enough tokens for other decoding methods, we set β=0.5 by default in our paper.

When α_1 is set to 0, it indicates that partial visible image features do not participate in the inference stage, and the performance lies between the original results and the best results. This suggests that the image features we designed provide richer image information but can introduce some noise in the absence of original image constraints.

The α_2 and α_3 respectively control the extent of image receptive fields and image spatial information. When either one acts alone, the overall performance does not reach optimal levels. However, optimal performance is achieved when both are utilized simultaneously.

D Case Study

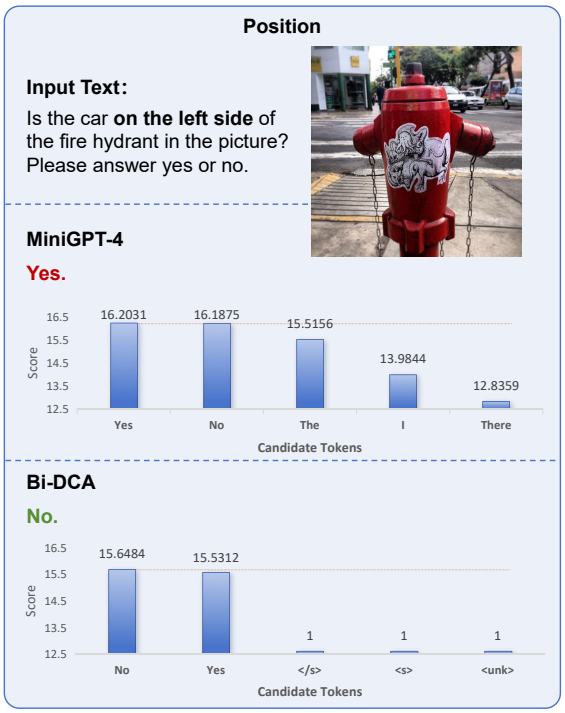

Figure 4: Comparison between greedy-based Bi-DCA and MiniGPT-4 on Position.

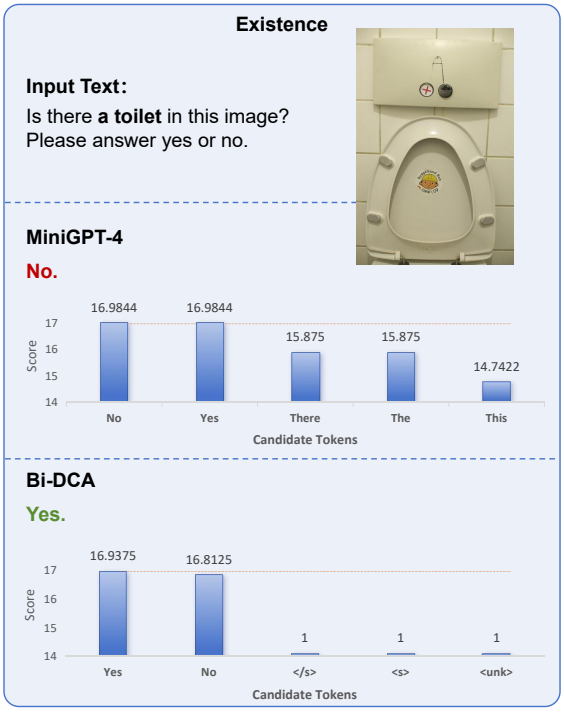

Figure 6: Comparison between greedy-based Bi-DCA and MiniGPT-4 on Existence.

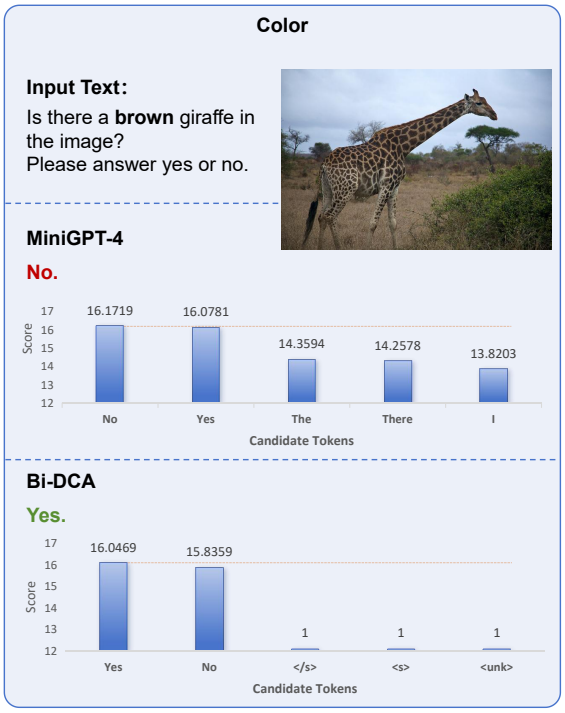

Figure 5: Comparison between greedy-based Bi-DCA and MiniGPT-4 on Color.

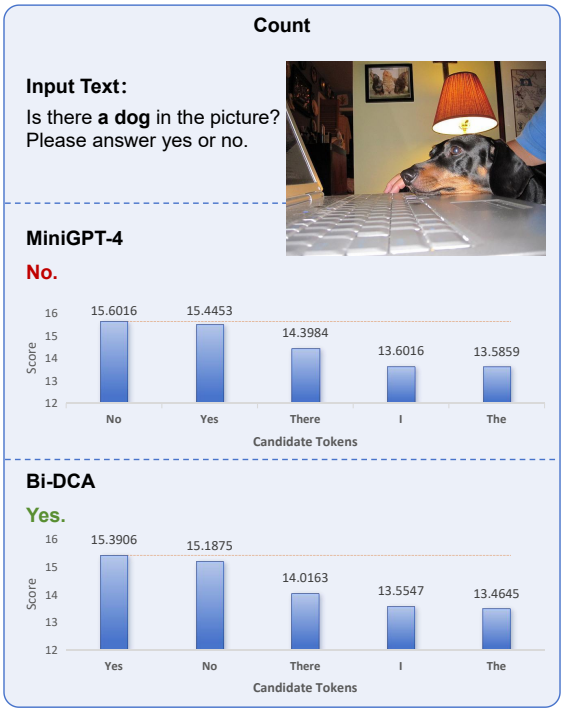

Figure 7: Comparison between beam-based Bi-DCA and MiniGPT-4 on Count.

Input Text:

Please describe this image in detail.

LLaVA-1.5:

In the image, a young man is playing with a **frisbee** in a fenced-in area. He is holding the **frisbee** in his hand and appears to be preparing to throw it. The fenced-in area seems to be a playground or a designated space for recreational activities.
There are a few other people in the background, but they are not the main focus of the scene. The young man with the frisbee is the center of attention as he enjoys his outdoor activity.

OPERA:

In the image, a young man is playing with a **frisbee** in a fenced-in area. He is holding the **frisbee** in his hand and appears to be preparing to throw it. The fenced-in area is likely a playground or a designated space for recreational activities.
There are a few other people present in the background, but they are not the main focus of the scene. The young man with the frisbee is the center of attention, enjoying his time outdoors and engaging in a fun activity.

Bi-DCA:

In the image, a young man is playing tennis on a court, holding a tennis racket and preparing to hit a tennis ball. He is wearing a striped shirt and appears to be focused on the game. The tennis ball is positioned close to one of his feet, while the racket is held in his other hand.
The tennis court is surrounded by a chain-link fence, providing a secure boundary for the game. The scene captures the essence of an outdoor recreational activity, showcasing the man's athleticism and dedication to the sport.

Input Text:

Please describe this image in detail.

LLaVA-1.5:

The image features a man standing in a bathroom, taking a selfie with his cell phone. He is holding the phone in front of his face, capturing a picture of himself. The man is wearing a green shirt and appears to be smiling as he takes the photo.

In the background, there is a **toilet** visible on the left side of the image, and a **sink** can be seen on the right side. **Additionally, there is a toothbrush placed near the sink, suggesting that the bathroom is well-equipped for daily hygiene routines.**

OPERA:

The image features a man standing in a bathroom, holding a cell phone in his hand. He is taking a selfie, capturing his reflection in the mirror. The man appears to be wearing a green shirt and has a beard.

The bathroom is equipped with a **sink** and a **toilet**. **The sink is located on the left side of the bathroom, while the toilet is situated on the right side. The man's reflection can be seen in the mirror, which is positioned above the sink.**

Bi-DCA:

The image features a man standing in a bathroom, holding a cell phone in his hand. He is taking a picture of himself in the mirror, capturing a selfie. The man is wearing a green shirt and appears to be focused on the task at hand.