# DAMRO: Dive into the Attention Mechanism of LVLM to Reduce Object Hallucination

**Xuan Gong, Tianshi Ming, Xinpeng Wang, Zhihua Wei***
Department of Computer Science and Technology, Tongji University
{2152095,2151569,wangxinpeng,zhihua_wei}@tongji.edu.cn

## Abstract

Despite the great success of Large Vision-Language Models (LVLMs), they inevitably suffer from hallucination. As we know, both the visual encoder and the Large Language Model (LLM) decoder in LVLMs are Transformer-based, allowing the model to extract visual information and generate text outputs via attention mechanisms. We find that the attention distribution of LLM decoder on image tokens is highly consistent with the visual encoder and both distributions tend to focus on particular background tokens rather than the referred objects in the image. We attribute to the unexpected attention distribution to an inherent flaw in the visual encoder itself, which misguides LLMs to over emphasize the redundant information and generate object hallucination. To address the issue, we propose DAMRO, a novel training-free strategy that **D**ive into **A**ttention **M**echanism of LVLM to **R**educe **O**bject Hallucination. Specifically, our approach employs classification token (CLS) of ViT to filter out high-attention outlier tokens scattered in the background and then eliminate their influence during decoding stage. We evaluate our method on LVLMs including LLaVA-1.5, LLaVA-NeXT and InstructBLIP, using various benchmarks such as POPE, CHAIR, MME and GPT-4V Aided Evaluation. The results demonstrate that our approach significantly reduces the impact of these outlier tokens, thus effectively alleviating the hallucination of LVLMs.

## 1 Introduction

Large Vision-Language Models (LVLMs) research (Dai et al., 2023; Liu et al., 2024b; Chen et al., 2023; Ye et al., 2023) has witnessed rapid advancement in the past few years, particularly demonstrating strong capabilities in visual reasoning tasks. However, LVLMs still face significant challenges related to object hallucination (Rohrbach et al., 2018), where the objects described in the generated text do not align with the visual ground truth of the input. This issue is prevalent across various models, posing a critical problem for the reliability and safety of LVLMs (Ahmad et al., 2023).

Recently, the issue of object hallucination in LVLMs has gained increasing attention. Early work has tried many methods, such as optimizing the training and fine-tuning methods (Sarkar et al., 2024; Xiao et al., 2024), incorporating external information or models, e.g. DETR (Carion et al., 2020)(Zhao et al., 2024; Chen et al., 2024), providing feedback on hallucinated information and reprocesses (Zhou et al., 2024; Yin et al., 2023). Efforts also include LLM decoding methods, like contrastive decoding (Leng et al., 2024; Favero et al., 2024) and other novel decoding methods (Huang et al., 2024).

These approaches mainly focus on improving the overall model architecture or specific modules within LVLMs, such as the visual encoder or LLM decoder. However, they often overlook the fundamental component of LVLMs, the Vision Transformer (ViT) structure (Dosovitskiy et al., 2021), and its impact on the hallucination generation mechanism during the LLM decoding stage.

Based on LLaVA-1.5 (Liu et al., 2024a), we explore the attention map in both the visual encoder and the LLM decoder. We find outlier tokens in the attention map of both components, which are highly consistent with each other. These high-norm outlier tokens often contain globally redundant visual information (Darcet et al., 2024). Additionally, our analysis reveals a correlation between attention to these tokens and the occurrence of object hallucination.

To address the aforementioned issue, we propose the **D**ive into **A**ttention **M**echanism of LVLM to **R**educe **O**bject Hallucination (DAMRO) method, as illustrated in Figure 1. DAMRO filters out high-norm outlier tokens from the ViT attention

*Corresponding author



*Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7696–7712
November 12-16, 2024 

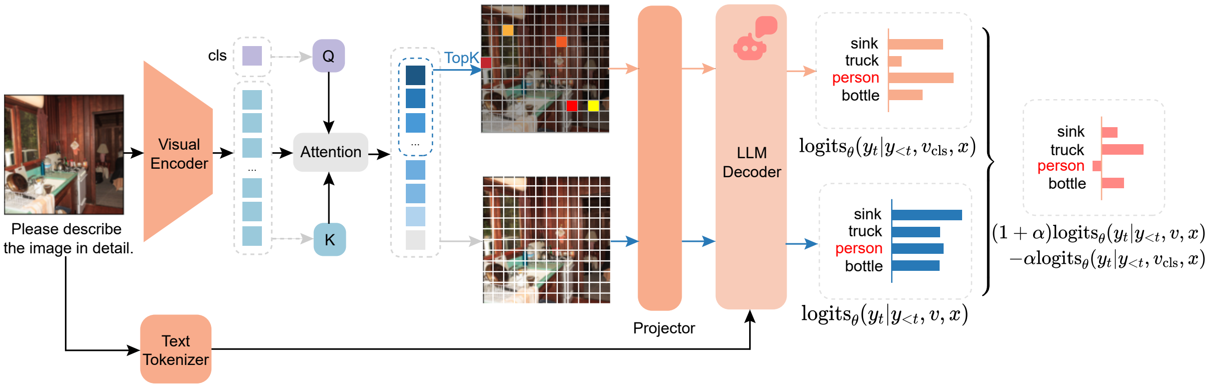


Figure 1: An overview of DAMRO. We utilize attention mechanism to filter the outlier tokens, and then apply contrastive decoding to mitigate the influence of outlier tokens in LLM decoding stage.

map, identifying them as negative tokens, and then projects them into the LLM along with normal tokens. Contrastive decoding is then applied to reduce the LLM decoder's reliance on these tokens that contain globally redundant information and to enhance its focus on object-level details, thus mitigating model hallucination.

Our method is training-free and does not introduce external information or models. It outperforms similar approaches such as M3ID (Favero et al., 2024) and VCD (Leng et al., 2024) in overall effectiveness. Additionally, since ViT is such a popular backbone of visual encoder (Yin et al., 2024) that our approach demonstrates strong generalizability due to its utilizing on attention mechanism.

In conclusion, our main contributions are summarized as follows:

- We conduct in-depth analysis of the relationship between the attention maps of the visual encoder and the LLM decoder, revealing a high consistency in the distribution of their outlier tokens.

- We analyze the impact of the consistency on object hallucination and design the DAMRO method to mitigate the hallucination in LVLMs.

- We demonstrate effectiveness of our method via extensive experiments on various models and benchmarks. Moreover, our training-free approach is applicable to most LVLMs without external knowledge or models.

## 2 Related Work

### 2.1 Hallucination in LVLMs

In LVLMs, hallucination refers to discrepancies between visual input (ground truth) and textual output. Hallucination is initially identified and studied in LLM research (Huang et al., 2023; Ji et al., 2023). However, LVLMs also suffer from hallucination, which is much more complex due to their intricate structure. Han et al. (2024) analyze hallucination from the perspective of training data bias. Tong et al. (2024), Jiang et al. (2024), and Huang et al. (2024) focus on structural causes, revealing the flaws in visual encoders, the misalignment of visual-textual modalities, and the inherent hallucinations of LLM respectively. Zhou et al. (2024) identify patterns in LVLM input and output, proposing object co-occurrence, model uncertainty, and the spatial positioning in sentence as causes. These studies reveal the mechanisms of hallucinations and offer new approaches to address this issue in LVLMs.

Unlike previous studies, we start by analyzing the attention maps of the visual encoder and LLM decoder, focusing on their distribution characteristics and correlations. This analysis provides new insights into object hallucination.

### 2.2 Contrastive Decoding to Mitigate Hallucination

Contrastive decoding (Li et al., 2023a) is first introduced in text generation tasks in LLMs to reduce noise by subtracting the distribution of an amateur model. To address hallucination issues in LVLMs, researchers have introduced contrastive decoding to improve model performance. Leng et al. (2024)

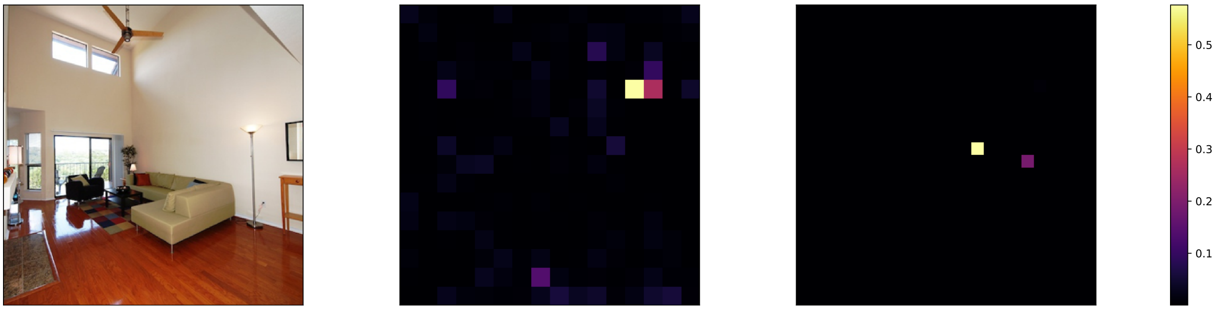


Figure 2: Attention map of visual encoder. **Left:** original image. **Middle:** attention map of InstructBLIP ViT (16x16). **Right:** attention map of LLaVA-1.5 ViT (24x24).

apply Gaussian noise to images to increase visual uncertainty. They use these noisy images as negative samples to subtract the LLM's prior and reduce object hallucination. Favero et al. (2024) employ pure text inputs as negative samples. They apply contrastive decoding to enhance the influence of visual information during text generation. Wang et al. (2024) introduce a disturbance instruction to force the model to output an error distribution, which is then subtracted to mitigate hallucination.

Given that our method draws on contrastive decoding and considering the generality and effectiveness of these methods, in section 5.1, we select VCD (Leng et al., 2024) and M3ID (Favero et al., 2024) as our baselines for experimental comparison.

# 3 Motivation

## 3.1 Problem Formulation

We segment the LVLM generation process into three distinct stages: Visual Encoding, Projection, and LLM Decoding. In the initial stage, an input image is divided into $n$ patches, each projected into a token embedding via Vision Transformer. The set of $n$ tokens is represented as $X_v = \{X_{v_i} | 0 \leq i < n\}$ . Then tokens are forwarded to the LLM after projection. Concurrently, the prompt is tokenized into tokens $X_l$ and is put into the LLM directly or indirectly.

In the decoding stage, we perform autoregressive decoding with the transformer, which is formulated in Eq. 1.

$$p_t = \text{softmax}(\text{logits}_\theta(y_t | y_{<t}, X_v, X_l)). \quad (1)$$

where $p_t$ represents probability distribution of next token $y_t$ in the $t$-th step of decoding, $y_{<t}$ represents the generated text from 0 to $t-1$ step and $\text{logits}_\theta$ represents the logit distribution. Then the LLM adopts a specific strategy to obtain the next token based on the probability distribution $p_t$.

We studied the impact of the visual token $X_v$ on $\text{logits}_\theta(y_t | y_{<t}, X_v, X_l)$ to reduce the likelihood of hallucination occurrence.

## 3.2 Drawbacks of ViT

The Vision Transformer (Dosovitskiy et al., 2021) has gained widespread favor as the backbone visual encoder for all LVLMs due to its superior visual representation capabilities. However, Darcet et al. (2024) find that there are always high-norm outlier tokens in ViT, which tend to appear in background regions with redundant patch information, containing minimal local information but a little global information.

The attention map of LVLMs' visual encoder also focus on a small number of high-norm outlier tokens, as illustrated in Figure 2. We posit that these outlier tokens embody the negative visual priors within the ViT. And when image tokens are projected and sent to the LLM, the LLM also tends to focus on these tokens due to their high attention value in visual encoder, leading to the ignorance of local information contained within other patches. This may result in a degradation of the model's fine-grained visual capabilities.

To validate the information contained within these tokens as perceived by the LLM, we conducted ablation experiments (results provided in Appendix B.3). The findings confirmed that these few tokens indeed contain substantial information, but are not accurate enough.

## 3.3 Outlier Tokens Cause Hallucination

Based on the aforementioned issues in ViT, we attempt to observe the attention maps of image tokens during LLM decoding stage. We find that LLM decoder attention map also features with a

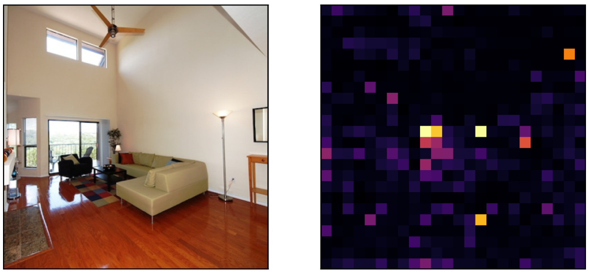

Figure 3: LLM decoder attention map of "plant" token (non-hallucinatory). It is evident that attention can accurately locate the position of the plotted plant.

Figure 4: LLM decoder attention map of "clock" token (hallucinatory). The attention mainly focus on the outlier tokens in the background, whose positions are the same in visual encoder attention map in the right sub-image of Figure 2.

few outlier tokens at the same position as visual encoder that get most of the attention compared to other tokens, as illustrated in Figure 5. We assume that this consistency is related to the occurrence of hallucination, where the LLM decoder pays more attention to outlier tokens identified in visual encoding stage. And we selected an example (Figure 3, 4) to demonstrate this correlation. To quantitatively characterize the consistency, we propose an evaluation metric $H_i$, where $S_v(i)$ denotes the set of top $i$ tokens of attention value from the visual encoder's attention map, while $S_l(i)$ represents the set of top $i$ tokens from the LLM decoder's attention map. And in this formulation, $|S|$ denotes the cardinality of the set $S$, which is the number of elements contained within $S$.

$$H_i = \frac{|S_v(i) \cap S_l(i)|}{i}. \tag{2}$$

We randomly select 1000 images from the val2014 subset in MSCOCO dataset (Lin et al., 2014) and query LLaVA-1.5 with the prompt "What can you see in this image ?" to get the descriptions from model. We use the generated captions and object words as two kinds of units and employed CHAIR (Rohrbach et al., 2018) to identify hallucinations. We then utilize metric $H_i$ to analyze the relation between the occurrence of hallucinations and the consistency of their distributions, as illustrated in Figure 6.

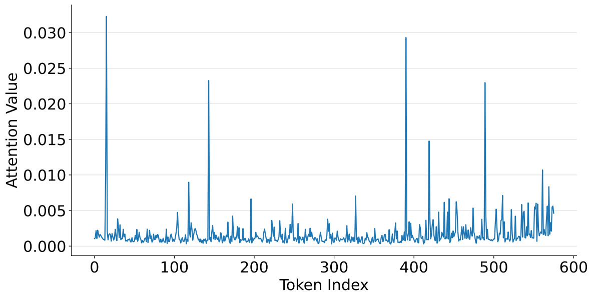


Figure 5: The proportion of the overall attention map in LLM decoder.

| Granularity | HA | Non-HA |
|---|---|---|
| sentence-level | 0.0554 | 0.0539 |
| object-level | 0.0605 | 0.0551 |

Table 1: $F$ Value results. HA: hallucinatory, Non-HA: non-hallucinatory. It is easily observed that at both the sentence level and the object level, the influence of outlier tokens from the visual encoder is greater when hallucinations occur.

Additionally, we found that the top three tokens with the highest attention score in the visual encoding stage accounted for more than 99% of the attention, as shown in Figure 7. To further verify the influence of these tokens, we analyzed the proportion of the same three tokens[1] in the attention map of LLM decoder. The evaluation metric of the influence is denoted as $F$, defined as

$$F = \frac{\sum_{j=1}^{3} ATT(L_v(j))}{\sum_{i=0}^{n-1} ATT(i)}. \tag{3}$$

where $L_v(i)$ represents the position of the token with $i$-th highest attention value in the visual encoder attention map and $ATT(i)$ represents the LLM decoder attention value of the token at position $i$.

Similarly, we use generated captions and object words as units to identify hallucinations. And we get the $F$ results in Table 1. It can be observed that outlier tokens in visual encoding stage indeed have influence on the subsequent LLM decoding stage, which is closely related to the occurrence of hallucinations.

[1]Unless otherwise specified, in this paper, the same tokens in the visual encoder and LLM decoder refer to tokens corresponding to the same spatial positions in the image.

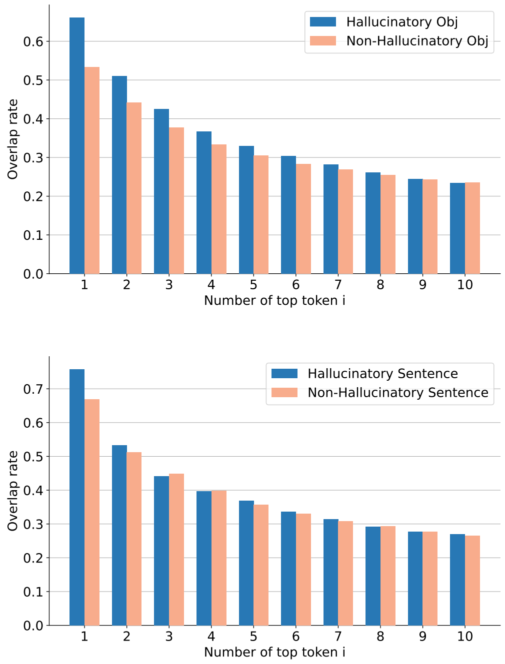


Figure 6: Top 1-10 outlier tokens overlap rate between visual encoder and LLM decoder. Both of object-level and sentence-level results show that hallucination tends to happen when overlap rate is higher, especially considering the top tokens.

# 4 Methods

## 4.1 Outlier Tokens Selection

In the final layer of self-attention in ViT, the class token [CLS] is generally used for classification (Dosovitskiy et al., 2021). The [CLS] token is used as the query vector in attention calculation with other visual tokens as key vector:

$$A_{\text{cls}} = \text{softmax}\left(\frac{Q_{\text{cls}}K^T}{\sqrt{d}}\right). \tag{4}$$

where $Q_{cls}$ is the result of the [CLS] token's query vector after being multiplied by the corresponding weights; $K^T$ is the result of all other image tokens' key vectors after being multiplied by their corresponding weights, and $d$ is the dimension of $Q_{cls}$.

We sample the top $k$ outlier tokens based on attention value between the class token [CLS] and spatial visual tokens, which is denoted as:

$$\text{token}_{\text{outlier}} = \underset{\text{token}_i}{\arg\max}(A_{\text{cls}}(\text{token}_i)). \tag{5}$$

For the selection of the top $k$, it is important to note that LLaVA-1.5 (Liu et al., 2024a) and InstructBLIP (Dai et al., 2023) have different ViT

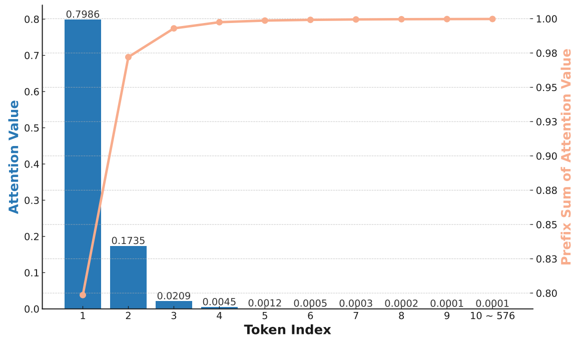


Figure 7: The proportion of the overall attention map occupied by tokens sorted by attention value in visual encoder.

structures. ViT in LLaVA-1.5 contains 576 (24x24) image tokens, whereas InstructBLIP has only 256 (16x16). The different numbers of image tokens lead to different choices in values of $k$ for the top $k$ selection. The difference in $k$ value will be discussed in detail in the ablation experiment in Appendix B.

## 4.2 Contrastive Decoding

We use Contrastive Decoding (Li et al., 2023a) to mitigate the impact of visual outlier tokens from the visual encoder on subsequent text generation. In LVLMs, Contrastive Decoding is typically conducted during the sampling process of LLM decoding, where the next token is determined based on the probability distribution in the logits space.

Answer generation in LLMs is an autoregressive process, in which the contrastive decoding is formulated as Eq. 6.

$$\begin{aligned} p_t = \text{softmax}((1+\alpha)&\text{logits}_\theta(y_t|y_{<t}, v, x) \\ &-\alpha\text{logits}_\theta(y_t|y_{<t}, v_{\text{cls}}, x)). \end{aligned} \tag{6}$$

where the probability distribution of the next token at step $t$ is $p_t$ with $x$ being the prompt input. $v_{\text{cls}} \in v$ is visual information filtered by [CLS] token from overall visual information $v$.

The probability distribution in the logits space attenuates the influence of previous outlier tokens on decoding. This allows the model to focus more on fine-grained semantic information and eliminates redundant information containing visual encoder priors, thus mitigating hallucinations in the LVLM.

To address the issue of excessive removal of global information, we introduced an adaptive plausibility constraint (Li et al., 2023a). In constrative decoding stage, we set a threshold $\beta$ to truncate the new probability distribution based on the confidence level of the original model's predictions. The specific form is shown in Eq. 7:

$$\mathcal{V}_{\text{head}}(y_{<t}) = \{y_t \in \mathcal{V} : p_\theta(y_t|v, x, y_{<t}) \geq \beta \max_w p_\theta(w|v, x, y_{<t})\}. \tag{7}$$

$\mathcal{V}_{\text{head}}$ serves as a filtering constraint for sampling the next token. The whole algorithm is further explained in Algo. 1.

**Algorithm 1** DAMRO

**Require:** text query $x$, image input $v$, visual encoder $I_\phi$.

Initialize empty output $y$ = [].
Large Language Model $\mathcal{M}_\theta$.
**for** t=0,1,2... **do**
  $I_\phi(v)_{i=1}^n \leftarrow \text{VisualEncoder}(v)$
  $\log p_{\text{origin}} \leftarrow \text{logits}_\theta(y_t|y_{<t}, I_\phi(v)_{i=1}^n, x)$
  $\text{Attn}_c^i \leftarrow \text{Attention}(\text{token}_{cls}, I_\phi(v)_{i=1}^n)$
  $I_{\text{outlier}} = \arg\max_I(\text{Attn}_c^i)$
  $\log p_{\text{negetive}} \leftarrow \text{logits}_\theta(y_t|y_{<t}, I_{\text{outlier}}, x)$
  Get token distribution in constrastive learning, $p_t \leftarrow \text{softmax}((1+\alpha)\log p_{\text{origin}} - \alpha \log p_{\text{negetive}})$,
  Considering adaptive plausibility constraint, $p_t = p_t$ if $p_t \geq \max(\log p_{\text{origin}})$ else 0
  Get next token using random sample strategy $y_t$.
  $y = [y, y_t]$
  **if** $y_t =$ <EOS> **then**
    break
  **end if**
**end for**
**return** Generated prompt $y$.

## 5 Experiments

### 5.1 Experimental Settings

**LVLM Models** We select three of the most representative LVLM models for evaluation: LLaVA-1.5-7b, LLaVA-NeXT-7b, and InstructBLIP-7b. For visual encoder, LLaVA-1.5 and LLaVA-NeXT share the same ViT backbone, both using ViT-L-336px pretrained from CLIP-L/14-336px (Radford et al., 2021). In contrast, InstructBLIP uses ViT-g/14 pretrained from EVA-CLIP (Sun et al., 2023). All three models use Vicuna[2] (Chiang et al., 2023) as the LLM module.

Regarding the connection module between the two modalities, LLaVA-1.5 and LLaVA-NeXT use MLP layers to bridge feature gap between vision and text modalities without changing the amount of image tokens in the LLM. Conversely, InstructBLIP employs Q-Former (Zhang et al., 2024) for modality alignment, which standardized the number of visual tokens in LLM to 32.

Our approach is based on LLaVA-1.5 in the analysis of Section 3.3. For more insights into generalizability, we also test our method on InstructBLIP, which has a significantly different structure compared to LLaVA-1.5, and we find that the performance still surpasses that of original model. This demonstrates that mitigating the impact of outlier tokens in the visual encoder is effective in alleviating hallucination across different projection modules.

**Baselines** We select two popular and training-free contrastive decoding methods: VCD (Leng et al., 2024) and M3ID (Favero et al., 2024). Both approaches aim to enhance the impact of visual features during the LLM decoding phase by eliminating language priors. VCD generates negative logits using Gaussian blurring, while M3ID generates negative logits using pure text that without visual information. Additionally, we include the original model for comparison to highlight the improvements over the baseline model. For detailed experimental hyperparameter settings of these baselines, please refer to Appendix A.

**Implementation Details** Considering the characteristics of different visual encoders, for LLaVA-1.5 and LLaVA-NeXT, we set $\alpha$ (Eq. 6) to 0.5 for CHAIR benchmark and 2 for other benchmarks and we select top 10 (Eq. 5) tokens as outlier tokens. For InstructBLIP, we set $\alpha$ to 1.5 for CHAIR benchmark and 0.5 for other benchmarks and we select top 4 tokens as outlier tokens. To avoid introducing additional factors, we directly use the probability distribution generated by the softmax function as the sampling distribution and employ the basic random sampling decoding strategy. For all experiments, the seed is set to 42, max_new_token is set to 1024 and $\beta$ (Eq. 7) is set to 0.1 .

### 5.2 Benchmarks and Experimental Results

**POPE** The Polling-based Object Probing Evaluation (POPE) (Li et al., 2023b) is a streamlined approach to assess object hallucination. LVLMs are required to respond to formatted questions in

[2]Vicuna-7b v1.5 for LLaVA-1.5 and LLaVA-NeXT, Vicuna-7b v1.1 for InstrutBLIP

| Base Model | Method | Precision | Recall | F1 Score | Accuracy |
|---|---|---|---|---|---|
| LLaVA-1.5 | Original | 88.63 | 73.76 | 80.48 | 82.08 |
| | VCD | 86.15 | **83.78** | **84.87** | 84.98 |
| | M3ID | **92.48** | 75.22 | 82.93 | 82.93 |
| | **DAMRO** | 88.84 | 81.09 | 84.72 | **85.31** |
| LLaVA-NeXT | Original | 92.28 | 75.58 | 83.07 | 84.57 |
| | VCD | 91.90 | 82.4 | 86.86 | 87.50 |
| | M3ID | **94.23** | 79.2 | 86.05 | 80.87 |
| | **DAMRO** | 90.02 | **85.40** | **87.60** | **87.87** |
| InstructBLIP | Original | 78.64 | 79.42 | 78.99 | 78.85 |
| | VCD | 84.88 | 79.93 | 81.96 | **82.56** |
| | M3ID | **90.59** | 70.58 | 79.33 | 81.60 |
| | **DAMRO** | 80.67 | **83.89** | **82.20** | 81.77 |

Table 2: Results of POPE. (The foundation model without methods is denoted as Original). The best value in the table is highlighted in **bold**, and the second best value is underlined.

the form: "Is there a <object> in the image?" with "Yes" or "No," . The answers to these questions alternate between "Yes" and "No," ensuring an equal 50% probability for each response. The complete POPE test is divided into three splits: random, popular and adversarial, in which missing objects are randomly selected, most frequently occurring in the dataset, and highly correlated with those present in the image respectively.

The dataset consists of 500 randomly selected images from the MSCOCO (Lin et al., 2014) validation set. To facilitate testing, we add the prompt "Please use one word to answer this question." to restrict LVLM responses to "Yes" or "No". Four key evaluation metrics are generated: Precision, Recall, F1 score, and Accuracy. We average the results across the three splits, and the outcomes are presented in Table 2. More details are shown in Appendix C.1.

**CHAIR** The Caption Hallucination Assessment with Image Relevance (CHAIR) (Rohrbach et al., 2018) is a widely used metric for evaluating object hallucination in image captioning tasks. CHAIR compares the captions generated by the LVLM with the ground truth to identify correctly and incorrectly described objects in the captions. It then calculates the proportion of objects mentioned in the captions that are not present in the images CHAIR evaluates hallucination on two dimensions: $\text{CHAIR}_S$ and $\text{CHAIR}_I$. The former calculates the proportion of sentences containing hallucinations at the sentence level, while the latter computes the proportion of hallucinated objects out of all identified objects at the object level. These two metrics can be formulated as follows:

| Model | Method | $C_S \downarrow$ | $C_I \downarrow$ |
|---|---|---|---|
| LLaVA-1.5 | Original | 12.4 | 7.2 |
| | VCD | 7.6 | 4.1 |
| | M3ID | 9.2 | 5.3 |
| | **DAMRO** | **6.0** | **3.6** |
| LLaVA-NeXT | Original | 4.2 | 9.0 |
| | VCD | **3.0** | **4.1** |
| | M3ID | 4.2 | 6.8 |
| | **DAMRO** | **3.0** | 5.2 |
| Instruct-BLIP | Original | 7.8 | 5.2 |
| | VCD | 3.2 | 1.9 |
| | M3ID | 5.2 | 3.7 |
| | **DAMRO** | **2.8** | **1.7** |

Table 3: Results of CHAIR. $C_S$: $\text{CHAIR}_S$, $C_I$: $\text{CHAIR}_I$.

$$\text{CHAIR}_S = \frac{|\{\text{captions w/ hallucinated objects}\}|}{|\{\text{all captions}\}|}.$$
$$\text{CHAIR}_I = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all mentioned objects}\}|}. \tag{8}$$

Similarly, we conducted the CHAIR evaluation on the MSCOCO dataset with 80 annotated object categories. We randomly selected 500 images from the validation set of COCO 2014 and used the prompt "Generate a short caption of this image." to obtain the generated captions.

The test results are shown in Table 3. It can

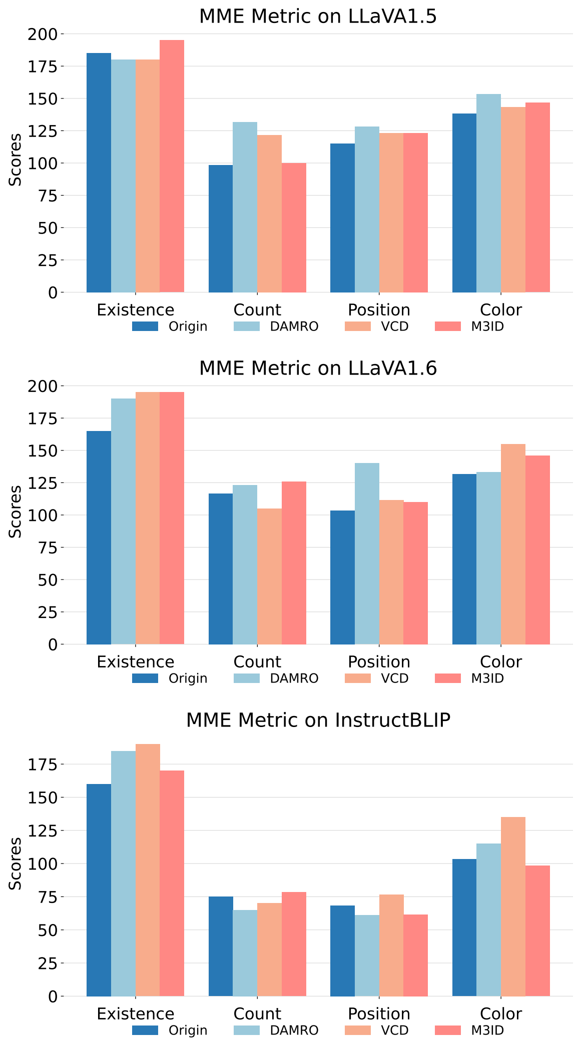


Figure 8: Results of MME.

be observed that, CHAIR scores on LLaVA-1.5 and InstructBLIP both surpassed the baseline compared to other methods, which achieve significant improvements in comparison with base model.

**MME Hallucination Subset** The Multimodal Large Language Model Evaluation (MME) (Fu et al., 2024) assesses LVLMs using a set of comprehensive metrics. Following the methodologies of Yin et al. (2023) and Leng et al. (2024), we adopted "existence" and "count" from the MME benchmark as object-level evaluation metrics, and "color" and "position" as attribute-level evaluation metrics. The experimental results in Figure 8 demonstrate that our approach generally improves performance across three models, confirming its effectiveness. However, for InstructBLIP, metrics for count and position show a decline. We hypothesize that this is due to the unique structure of InstructBLIP, which relies on certain outlier tokens for spatial reasoning. Compared to the LLaVA series of foundation models, InstructBLIP has significantly weaker positional capabilities, possibly explaining the reduced effectiveness of our approach for this model. Experiment Details are shown in Appendix C.2.

| Model | Method | A | D |
|---|---|---|---|
| LLaVA-1.5 | Original | 5.356 | 5.067 |
| | **DAMRO** | **6.611** | **6.078** |
| LLaVA-NeXT | Original | 6.456 | 6.332 |
| | **DAMRO** | **7.189** | **6.656** |
| InstructBLIP | Original | 5.833 | 5.400 |
| | **DAMRO** | **6.756** | **5.967** |

Table 4: Results of GPT4V-aided evaluation. A: accuracy, D: detailedness.

**GPT4V-Aided Evaluation** The GPT-4V-aided evaluation employs GPT-4V[3] as an evaluator to compare the outputs of two LVLM assistants. GPT-4V assigns scores out of 10 based on two criteria: 1) accuracy, which measures how accurately each assistant describes the image, and 2) detailedness, which evaluates the richness of necessary details in the responses. We select LLaVA-QA90 [4] for our tests on GPT-4V. The dataset consists of 30 images from COCO val2014, each paired with 3 questions to comprehensively evaluate the capabilities of LVLMs. Table 6 presents the overall scores of GPT-4V in terms of accuracy and detailedness, with detailed results provided in the appendix C.3.

## 6 Conclusions

In this paper, we investigate the relationship between the attention maps of the visual encoder and the LLM decoder, and explore its impact on the mechanism of object hallucination in LVLMs. Based on our analysis of attention mechanism, we propose the Dive into Attention Mechanism to mitigate object hallucination (DAMRO) method. Our method demonstrates its effectiveness and generalizability on various models and benchmarks. Experiments show that our method effectively reduces hallucination issues in LVLMs across multiple domains, especially in fine-grained semantic hallucinations. Additionally, we hope our findings on Encoder-Decoder attention mechanism will inspire

[3] https://openai.com/index/gpt-4v-system-card/

[4] https://github.com/haotian-liu/LLaVA/blob/main/playground/data/coco2014_val_gpt4_qa_30x3.jsonl

further research on LVLM foundation model structures.

## Limitations

Our method (DAMRO) is based on the relationship between the attention mechanisms of the visual encoder and the LLM decoder. It relies solely on empirical analysis and lacks further theoretical proof. Additionally, we have not conducted a detailed exploration of more complex projection modules in the visual encoder and LLM decoder (e.g. QFormer (Zhang et al., 2024)). With the rapid development and continual refinement of LVLM models, whether our method remains applicable to future models poses a significant challenge.

## Acknowledgements

The work is partially supported by the National Nature Science Foundation of China (No. 62376199, 62076184, 62076182) and Shanghai Science and Technology Plan Project (No.21DZ1204800).

## References

Muhammad Aurangzeb Ahmad, Ilker Yaramis, and Taposh Dutta Roy. 2023. Creating trustworthy llms: Dealing with hallucinations in healthcare ai. *Preprint*, arXiv:2311.01463.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020*, pages 213–229, Cham. Springer International Publishing.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechu Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *ArXiv preprint*, abs/2310.09478.

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

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. `https://vicuna.lmsys.org`. Accessed: 2024-06-13.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. 2024. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

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

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

Tianyang Han, Qing Lian, Rui Pan, Renjie Pi, Jipeng Zhang, Shizhe Diao, Yong Lin, and Tong Zhang. 2024. The instinctive bias: Spurious images lead to hallucination in mllms. *Preprint*, arXiv:2402.03757.

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

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13418–13427.

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

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

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

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023a. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312, Toronto, Canada. Association for Computational Linguistics.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.

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*, pages 740–755, Cham. Springer International Publishing.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium. Association for Computational Linguistics.

Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Ö. Arık, and Tomas Pfister. 2024. Mitigating object hallucination via data augmented contrastive tuning. *Preprint*, arXiv:2405.18654.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *Preprint*, arXiv:2303.15389.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9568–9578.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15840–15853, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Wenyi Xiao, Ziwei Huang, Leilei Gan, Wanggui He, Haoyuan Li, Zhelun Yu, Hao Jiang, Fei Wu, and Linchao Zhu. 2024. Detecting and mitigating hallucination in large vision language models via fine-grained ai feedback. *Preprint*, arXiv:2404.14233.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Mingshi Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. mplug-owi2: Revolutionizing multi-modal large language model with modality collaboration. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13040–13051.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *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. 2023. Woodpecker: Hallucination correction for multimodal large language models. *Preprint*, arXiv:2310.16045.

Qiming Zhang, Jing Zhang, Yufei Xu, and Dacheng Tao. 2024. Vision transformer with quadrangle attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3608–3624.

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

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024. Analyzing and mitigating object hallucination in large vision-language models. In *The Twelfth International Conference on Learning Representations*.

## A More Implementation Details

For the baselines M3ID and VCD, we employ the same direct sampling strategy as DAMRO. Throughout the entire experiment, our experimental hyperparameters remain consistent. The hyperparameters are listed in the table below:

| Hyperparameters | Value |
|---|---|
| Forgetting Factor(POPE) $\gamma$ | 0.2 |
| Forgetting Factor(CHAIR, MME) $\gamma$ | 0.01 |
| Threshold | 0.9 |

Table 5: M3ID Hyperparameters Settings.

| Hyperparameters | Value |
|---|---|
| Amplification Factor $\alpha$ | 1 |
| Adaptive Plausibility Threshold $\beta$ | 0.1 |
| Diffusion Noise Step | 999 |

Table 6: VCD Hyperparameters Settings.

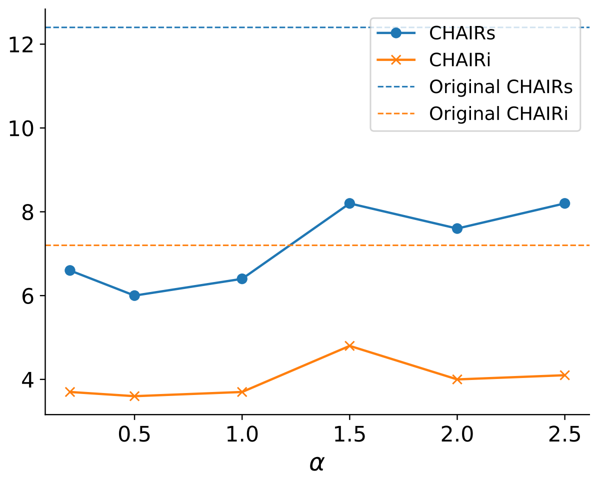


Figure 9: Ablation study of $\alpha$ in LLaVA-1.5, top k=10.

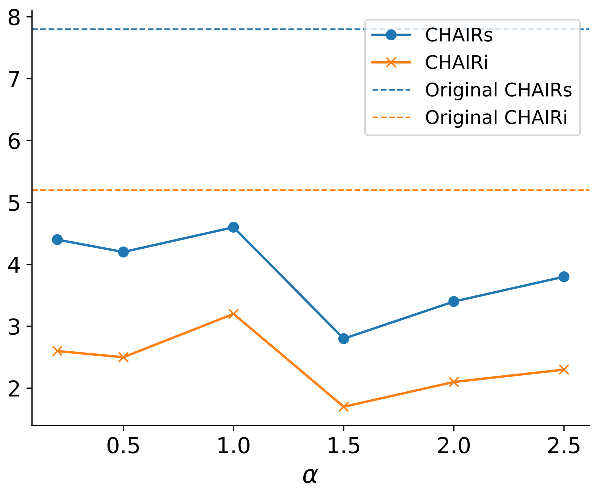


Figure 10: Ablation study of $\alpha$ in InstructBLIP, top k=4.

## B Ablation Study

Considering that CHAIR can more precisely assess the generative capabilities of the model, and given that LLaVA-1.5 and LLaVA-NeXT have similar model structures, we choose to test the parameter sensitivity of DAMRO on LLaVA-1.5 and Instruct-BLIP using CHAIR. The following two parameter ablation experiments are based on this setup. As for how many visual tokens are enough, we conduct ablation experiments on LLaVA-1.5 using POPE, CHAIR and MME benchmarks.

### B.1 Effect of $\alpha$ in Visual Contrastive Decoding

The results of the experiments with LLaVA-1.5 and InstructBLIP are shown in Figure 9 and Figure 10. It can be observed that when the value of $\alpha$ is too large or too small, the performance of the models deteriorates. $\alpha$ highlights the adjustment strength for outliers in our method, and the optimal adjustment strength varies for different models.

### B.2 Effect of Outlier Token Number top $k$

We use hyperparameters to define the number of outlier tokens, which vary across different visual encoders. Removing the top k outlier tokens aims to eliminate the redundant negative information they carry. However, this redundant information also contains a certain degree of global information, which can be beneficial for the results. Therefore, it is crucial to reasonably select the top k for our method. The results of the ablation experiments are shown in Figure 11 and Figure 12.

### B.3 How Many Visual Tokens are Enough

We conduct experiments using LLaVA-1.5 on CHAIR, POPE(only on random split), and MME, and found that a small number of visual tokens, or even a single token, can contain the basic information of an entire image. POPE,CHAIR,MME results are shown in Table 7,Table 8 and Table 9 respectively. Additionally, we select some images and examples from these CHAIR experiments, as shown in Figure 13 and Figure 14. It is evident that a few tokens indeed contain a large amount of information. However, the error rate of this information is quite high, easily leading to the co-occurrence of related objects, which reflects the priors of the visual encoder.

An interesting phenomenon is that using only a small number of tokens, some metric results are actually better than using more tokens. We attribute this to the fact that the LLM's attention to visual tokens cannot accurately capture the information they contain. Therefore, this also provides an idea for better selection and acquisition of effective tokens in future LVLM models.

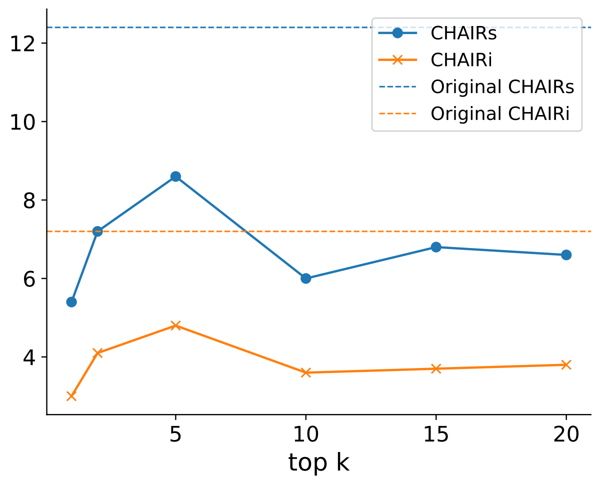


Figure 11: Ablation study of top $k$ in LLaVA-1.5, $\alpha$=0.5.

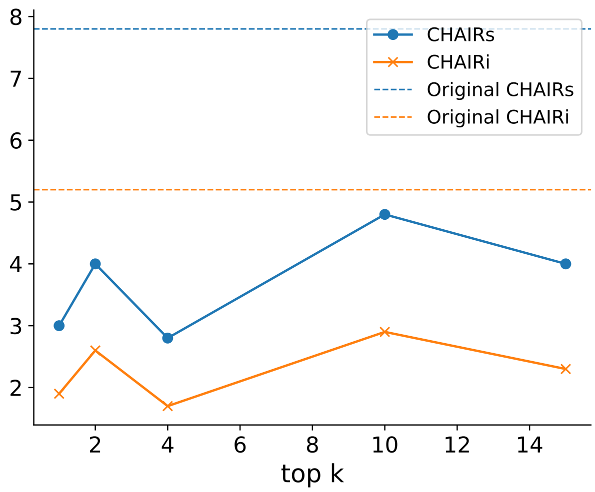


Figure 12: Ablation study of top $k$ in InstructBLIP, $\alpha$=1.5.

| | Precision | Recall | F1 | Accuracy |
|---|---|---|---|---|
| top 1 | 89.93 | 70.87 | 79.27 | 81.47 |
| top 2 | 90.47 | 69.60 | 78.67 | 81.13 |
| top 5 | 93.29 | 64.93 | 76.57 | 80.13 |
| top 10 | 94.76 | 63.93 | 76.35 | 80.20 |
| top 100 | 95.50 | 66.47 | 78.38 | 81.67 |
| **all** | **92.32** | **73.73** | **81.97** | **83.80** |

Table 7: POPE results with token numbers changed.

# C Detailed Results on POPE, MME and GPT4V-Aided Evaluation

## C.1 POPE Details

The detailed results of POPE on different sub-datasets are shown in Table 10.Our method achieved excellent results across different subsets.

| | CHAIRs ↓ | CHAIRi ↓ | Recall ↑ |
|---|---|---|---|
| top1 | 58.6 | 18.4 | 61.4 |
| top2 | 53.6 | 17.0 | 61.0 |
| top5 | 57.8 | 15.1 | 67.0 |
| top10 | 50.6 | 14.4 | 60.5 |
| top100 | 57.8 | 15.1 | 67.0 |
| **all** | **60.2** | **16.8** | **68.1** |

Table 8: CHAIR results with token numbers changed.

## C.2 MME Details

The detailed results of MME are shown in Table 11

## C.3 GPT4V-aided Evaluation Details

To evaluate open-ended generation, we utilize GPT-4V to assess the accuracy and detailedness of LVLMs' responses. The specific configurations are detailed in Table 12. Additionally, two illustrative evaluation cases are presented in Figure 15 and Figure 16.

| | **existence** | **count** | **position** | **color** | **total** |
|---|---|---|---|---|---|
| top1 | 175.00 | 81.67 | 98.33 | 116.67 | 471.67 |
| top2 | 180.00 | 91.67 | 96.66 | 136.66 | 504.99 |
| top5 | 168.33 | 90.00 | 116.67 | 125.00 | 500.00 |
| top10 | 178.33 | 80.00 | 96.66 | 118.33 | 473.32 |
| top100 | 170.00 | 80.00 | 90.00 | 121.67 | 461.67 |
| **all** | **185.00** | **98.30** | **115.00** | **138.30** | **536.30** |

Table 9: MME results with token numbers changed.

| **Model** | **Dataset** | **Method** | **Precision** | **Recall** | **F1** | **Accuracy** |
|---|---|---|---|---|---|---|
| LLaVA-1.5 | random | Original | 92.321 | 73.733 | 81.987 | 83.800 |
| | | **DAMRO** | 94.557 | 81.067 | 87.294 | **88.200** |
| | | VCD | 91.886 | **83.8** | **87.657** | **88.200** |
| | | M3ID | **96.331** | 75.267 | 84.506 | 86.200 |
| | popular | Original | 89.700 | 73.733 | 80.937 | 82.633 |
| | | **DAMRO** | 89.280 | 81.067 | 84.976 | 85.667 |
| | | VCD | 87.231 | **83.800** | **85.481** | **85.767** |
| | | M3ID | **92.923** | 75.267 | 83.168 | 84.767 |
| | adversarial | Original | 83.864 | 73.800 | 78.511 | 79.800 |
| | | **DAMRO** | 82.677 | 81.133 | **81.898** | **82.067** |
| | | VCD | 79.343 | **83.733** | 81.479 | 80.967 |
| | | M3ID | **88.185** | 75.133 | 81.138 | 82.533 |
| LLaVA-NeXT | random | Original | 96.500 | 75.600 | 84.785 | 86.433 |
| | | **DAMRO** | 94.749 | **85.400** | **89.832** | **90.333** |
| | | VCD | 96.187 | 82.400 | 88.760 | 89.567 |
| | | M3ID | **97.457** | 79.200 | 87.385 | 88.567 |
| | popular | Original | 92.571 | 75.600 | 83.229 | 84.767 |
| | | **DAMRO** | 90.594 | **85.400** | **87.920** | **88.267** |
| | | VCD | 92.170 | 82.400 | 87.010 | 87.700 |
| | | M3ID | **93.913** | 79.200 | 85.931 | 87.033 |
| | adversarial | Original | 87.761 | 75.533 | 81.189 | 82.500 |
| | | **DAMRO** | 84.720 | **85.400** | **85.059** | 85.000 |
| | | VCD | 87.340 | 82.400 | 84.803 | 85.233 |
| | | M3ID | **91.314** | 79.200 | 84.827 | **85.833** |
| InstrucBLIP | random | Original | 81.975 | 79.133 | 80.523 | 80.867 |
| | | **DAMRO** | 85.890 | **84.000** | **84.934** | 85.100 |
| | | VCD | 89.694 | 80.067 | 84.607 | **85.433** |
| | | M3ID | **93.451** | 70.400 | 80.304 | 82.733 |
| | popular | Original | 79.112 | 79.067 | 79.093 | 79.100 |
| | | **DAMRO** | 80.089 | **83.667** | **81.839** | 81.433 |
| | | VCD | 83.907 | 79.600 | 81.697 | **82.167** |
| | | M3ID | **90.000** | 70.800 | 79.254 | 81.467 |
| | adversarial | Original | 74.829 | 80.067 | 77.359 | 76.567 |
| | | **DAMRO** | 76.010 | **84.067** | **79.835** | 78.767 |
| | | VCD | 81.052 | 80.133 | 79.59 | **80.700** |
| | | M3ID | **88.314** | 70.533 | 78.428 | 80.600 |

Table 10: Detailed results of POPE on different sub-datasets.

| Model | Method | Object-level | | Attribute-level | | Total Scores |
|---|---|---|---|---|---|---|
| | | Existence | Count | Position | Color | |
| LLaVA-1.5 | Original | 185.00 | 98.30 | 115.00 | 138.30 | 536.60 |
| | VCD | **195.00** | 100.00 | 123.33 | 146.67 | 565.00 |
| | M3ID | 180.00 | 121.67 | 123.33 | 143.33 | 568.33 |
| | **DAMRO** | 180.00 | **131.67** | **128.30** | **153.30** | **593.27** |
| LLaVA-NeXT | Original | 165.00 | 116.67 | 103.33 | 131.66 | 516.66 |
| | VCD | **195.00** | **126.00** | 110.00 | 146.00 | 577.00 |
| | M3ID | **195.00** | 105.00 | 111.67 | **155.00** | 566.67 |
| | **DAMRO** | 190.00 | 123.33 | **140.00** | 133.33 | **586.66** |
| InstructBLIP | Original | 160.00 | 75.00 | 68.30 | 103.3 | 406.60 |
| | VCD | 170.00 | **78.30** | 61.67 | 98.33 | 408.30 |
| | M3ID | **190.00** | 70.00 | **76.67** | **135.00** | **471.67** |
| | **DAMRO** | 185.00 | 65.00 | 61.00 | 115.00 | 426.00 |

Table 11: Detailed results of MME.


Figure 13: A case illustrates the generative ability of tokens. We use the prompt "Please describe this image in detail." to get answers from different. Hallucinated words are marked in red.


Figure 14: A case illustrates the generative ability of tokens. We use the prompt"Please describe this image in detail." to get the answers. Hallucinated words are marked in red.

| **GPT-4V(ision) Prompt** |
|---|
| 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:<br>1: Accuracy: whether the response is accurate with respect to the image content. Responses with fewer hallucinations should be given higher scores.<br>2: Detailedness: whether the response is rich in necessary details. Note that hallucinated descriptions should not count as necessary details.<br>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.<br>[Assistant 1]<br>{}<br>[End of Assistant 1]<br><br>[Assistant 2]<br>{}<br>[End of Assistant 2]<br><br>Output format:<br>Accuracy:<br>Reason:<br>Detailedness:<br>Reason: |

Table 12: The prompt used for GPT-4V(ision) evaluation.

[Input Image]

[Input Instruction] What is the man doing in the image?

[Model Reponse]

InstructBLIP

The man in the image is holding up various items, including a suit jacket, a husband coat, a man suit, a robe, and a doll suit. He is possibly preparing to pavck or hang these items up in his residence.

Ours

The man in the image is holding up two garment covers, one over a suitcase and the other over a dress.

GPT-4 Evaluation

Accuracy: **1 8**

Reason: Assistant 1 incorrectly describes multiple items that are not present in the image, such as a suit jacket, a husband coat, a man suit, a robe, and a doll suit, leading to a high level of hallucination. In contrast, Assistant 2 accurately describes the man holding up two garment covers, which is consistent with the image content. There is a minor error in stating that one garment cover is over a suitcase and the other over a dress; it's actually over a blue bag, not necessarily a suitcase.

Detailedness: **2 5**

Reason: Assistant 1 provides a more detailed (though inaccurate) description by listing various clothing items but does so incorrectly, causing hallucination. Meanwhile, Assistant 2's description is concise and relatively accurate but lacks in-depth detail. Hence, while Assistant 2's description is more accurate, it is also less detailed.

Figure 15: DAMRO's performance on reducing hallucinations on on InstructBLIP.

[Input Image]

[Input Instruction] Describe the following image.

[Model Reponse]

LLaVA1.5

The scene is set in a hallway with a luggage cart, heaped with clothes, bags and suitcases. There's an abundance of suitcases of varying sizes placed on top of the cart. Two handbags can be seen among the suitcases as well. Several people are present in the vicinity, one of whom is positioned near the r ight side of the hallway, while others are further in the background. A TV is visible on the lower left side, partially hidden by the luggage cart and suitcases. Various chairs are placed in different spots of the hall, suggesting a social or work environment.

Ours

The image shows a luggage cart in a lobby filled with various suitcases and bags, waiting for passengers to pick up their belongings. The cart is packed with multiple suitcases and backpacks of different sizes and shapes. Several people can be seen around the lobby, with some standing closer to the cart, while others are further away. In the background, there are TVs mounted on the wall, indicating the location might be a hotel lobby or a waiting area for travelers. A handbag is also placed on the cart among the suitcases.

GPT-4 Evaluation

Accuracy: **4 5**
Reason: Assistant 1 mentioned several errors such as the presence of a TV and various chairs which are not visible in the image. Furthermore, the abundance of suitcases and two handbags are also hallucinations. Assistant 2 made fewer errors, mainly misrepresenting the number and type of bags, and imagining mounted TVs, which were not present.

Detailedness: **6 7**
Reason: Assistant 1 provided more details about the setting, such as mentioning the hallway and the chairs. Despite the hallucinations, the description also covered many elements of the scene. Assistant 2 was relatively detailed, mentioning the luggage cart and suggesting a hotel lobby or waiting area. The accuracy contributed to the slightly higher score for Assistant 2 in detailedness.

Figure 16: DAMRO's performance on reducing hallucinations on LLaVA-1.5-7b.