MaskCD: Mitigating LVLM Hallucinations by Image Head Masked Contrastive Decoding

Anonymous ACL submission

Abstract

Large vision-language models (LVLMs) have shown remarkable performance in visual-language understanding for downstream multimodal tasks. While their capabilities are improving, problems emerge simultaneously. Among those problems, the hallucinations have attracted much attention, which stands for the phenomenon where LVLMs generate contradictory content to their input visual and text contents. Many approaches have been proposed to deal with this issue, such as contrastive decoding and attention manipulation. However, contrastive decoding methods struggle in constructing appropriate contrastive samples, and attention manipulation methods are highly sensitive, lacking stability. In this work, we propose image head **Mask**ed **C**ontrastive **D**ecoding (**MaskCD**). Our approach utilizes the "image heads" in LVLMs, masking them to construct contrastive samples for contrastive decoding. We evaluated MaskCD on LLaVA-1.5-7b and Qwen-VL-7b, using various benchmarks such as CHAIR, POPE, and MME. The results demonstrate that MaskCD effectively alleviates the phenomenon of hallucinations and retains the general capabilities of LVLMs.

1 Introduction

Large Language Models (LLMs) (Brown et al., 2020,Brown et al., 2020,OpenAI, 2023) have achieved remarkable success in understanding human instructions and performing diverse tasks. Building on this progress, recent efforts have extended LLMs to develop Large Vision-Language Models (LVLMs) (Bai et al., 2023b,Li et al., 2023a,Dai et al., 2023,Zhu et al., 2023,Ye et al., 2023,Liu et al., 2023b), which integrate visual and textual modalities for multimodal reasoning. Although researchers have already achieved remarkable success in applying LVLMs into several tasks, problems have emerged as well. Within these problems, the hallucination (Zhang et al., 2024,Kamath et al., 2023,Li et al., 2023b) has attracted significant attention.

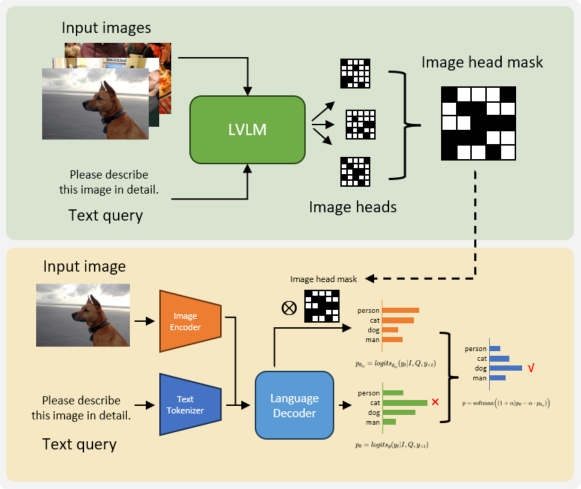
Input images
LVLM
Please describe this image in detail.
Text query
Image heads
Image head mask
Input image
Image Encoder
Text Tokenizer
Language Decoder
Image head mask
⊗
person
cat
dog
man
$p_{\theta_m} = logits_{\theta_m}(y_t|I, Q, y_{<t})$
$p_\theta = logits_\theta(y_t|I, Q, y_{<t})$
×
√
$p = \text{softmax}((1+\alpha)p_\theta - \alpha \cdot p_{\theta_m}))$

Figure 1: **Pipeline of MaskCD.**The upper part shows the first step. The image head mask is constructed by querying LVLM with images and prompt texts. Then, the lower part shows how to use the image head mask in the process of contrastive decoding.

The hallucination of LVLMs is a phenomenon in which models tend to generate contradictory contents for the inputs, especially images. This may manifest as generating non-existent objects, mistakenly described attributes, or non-sense sentences. All kinds of hallucinations enormously lower users' trust in the model and even cause fatal damage when applied in real-world tasks like auto-driving and medical image processing.

To mitigating the hallucination phenomenon, researchers have promoted multiple methods, which could be classified into two main categories according to whether training is needed. Firstly, training-involved methods (Lee et al., 2024,Chen et al., 2024,Liu et al., 2024a) collect massive delicately-constructed data to fine-tune or post-train the LVLMs, so as to teach models to generate less hallucinated content. It features with a high-

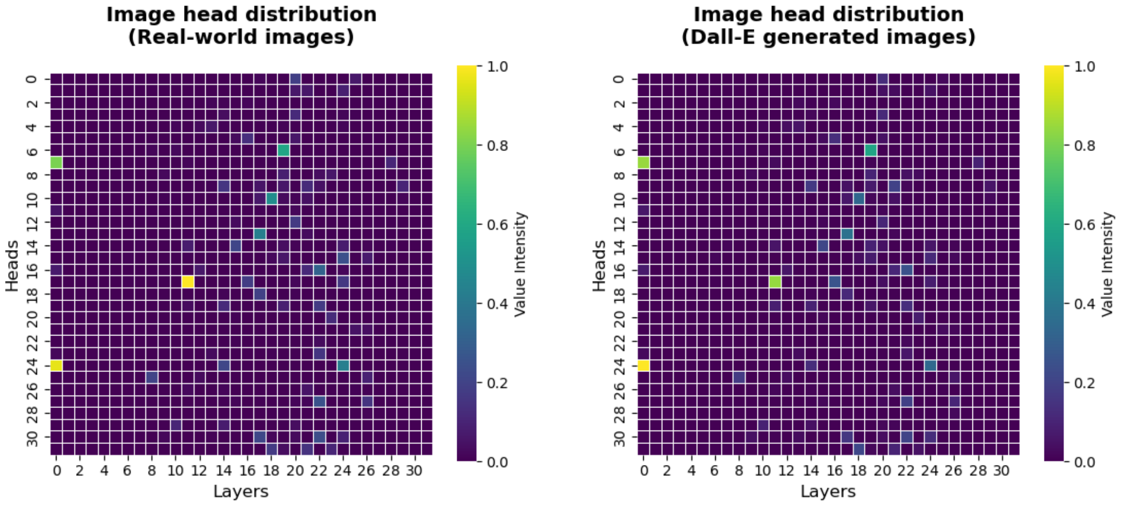

Image head distribution (Real-world images)

Image head distribution (Dall-E generated images)

Figure 2: **Visualization of image heads in LLaVA-1.5-7b.** The left figure shows the image head distribution of real-world images, while the right one represents the results of Dall-E generated artificial images. It is evident that there are certain heads that tend to pay high attention on image tokens, therefore we name them with "image head".

cost of computation resources and massive human labor. On the contrary, training-free methods are developed to alleviate hallucination at a lower cost. Prevailing methods include contrastive decoding (CD) (Leng et al., 2024,Favero et al., 2024,Woo et al., 2024) and attention manipulation (Tu et al., 2025,Huang et al., 2024,Liu et al., 2024b). CD methods need an injured input as the bad sample, whose output logits will be subtract from the original one. It would take two times of inference cost, one for the original input, the other for the injured one. But with delicately constructed bad samples, CD methods have presented prominent performance in mitigating the LVLMs' hallucination phenomenon. Recently, with the progress in understanding models' inner working mechanisms, methods of attention manipulation have come up. Specifically, abnormal attention map phenomena, such as excessive local attention in the attention sink phenomenon (Favero et al., 2024,Xiao et al., 2024), will be directly reduced or redistributed. This method features in better align visual and text modal, enabling models to better and truly utilize visual information.

Although the need for training has been eliminated, both the CD and attention manipulation methods have their drawbacks. The performance of CD methods is heavily depended on the quality of the constructed bad sample. If the injured sample still contains a lot of useful information, then the contrast operation may even cause worse results. For the attention manipulation methods, models are highly sensitive to changes in the attention score and are not as stable as CD methods in terms of overall testing scenarios.

Therefore, we hope to construct high-quality bad samples through the angle of attention distribution, thereby combining the advantages of CD methods and attention manipulation methods, balancing stability and hallucination-mitigating performance. To observe the model's attention preference to images at head-level, we randomly select 500 images in the validation set of COCO 2014(Lin et al., 2014) and 500 Dall-E generated artificial images from MMrel (Nie et al., 2024). Put them into LLaVA-1.5-7b (Liu et al., 2023b) with the prompt text "Please describe this image in detail." and record the sum of the attention scores obtained by each head in each layer of the model for each token generation. Finally, under different thresholds, the number of times each head pays excessive attention to the image token during the generation of each token is calculated.

The normalized result is visualized in Figure 2. We observed that whether real-world images or AI-generated images, there are certain heads in LLaVA-1.5-7b that prefer to give image tokens comparably high attention scores. Since they present an inclined focus on visual information, we name them "image heads". Given that the essence of CD methods is making the subtracted samples contain only invalid information as much as possible, we choose to mask these image heads to construct bad samples, so as to prevent the bad samples from accessing useful visual information more precisely.

In this way, we proposed image head Masked Contrast Decoding (MaskCD), which uses image head attention masks to construct delicate bad samples and attain significant hallucination-mitigating performance.

Our contributions can be summarized as follows:

- We identify "image heads" in LVLMs that disproportionately attend to image tokens.
- We propose MaskCD, a contrastive decoding method that features in using image head masking to construct degraded visual inputs.
- We demonstrate, through extensive experiments, that MaskCD outperforms existing hallucination mitigation methods across multiple benchmarks while preserving general model capabilities.

2 Related Work

2.1 Large Vision-Language Model

Recently, efforts have been made to enhance Large Vision-Language Models, aiming to equip LLMs with the ability to process visual information like images or videos. LVLMs are typically constructed by three components: a visual encoder to extract visual features, a modality connection module to bridge visual and text modal, and an LLM for further tasks. The visual encoder and LLM are typically fixed pretrained models; common choices are CLIP model (Radford et al., 2021) variants for the visual encoder, and LLaMA (Touvron et al., 2023) or Vicuna (Chiang et al., 2023) for the LLM.

Research focuses on optimizing modality connection modules, so as to better utilize visual and text information at the same time. Different connection modules lead to diffent LVLM types: cross-attention module in Flamingo(Alayrac et al., 2022), Q-former in BLIP-2(Li et al., 2023a), and simple linear layer in LLaVA(Liu et al., 2023b) model series.

2.2 Hallucination in LVLMs

Multimodal hallucination phenomenon, typically presented as LVLM generates inconsistent content from the input, especially those it contradictory with visual information. For example, in the image captioning task, LVLM may generate objects that do not exist in the input images (Li et al., 2023b), or mistakenly describe attribution of existing objects like counts, color and spatial relationship (Kamath et al., 2023).

The methods for alleviating LVLM hallucinations can be classified according to whether training is required. Training-involved methods typically uses constructed data to fine-tune or post-train LVLMs. For example, Hu et al. (2023), Liu et al. (2023a) uses contrastive question-answer pairs to fine-tune LVLMs, and Sun et al. (2024) employs Reinforcement Learning from Human Feedback (RLHF) to enhance multimodal connections. Training-free methods are prevailed by contrastive decoding and attention manipulation. The core of CD methods is constructing bad samples that contain useful information as less as possible. Different constructing means like image editing(Woo et al., 2024,Leng et al., 2024), text editing(Wang et al., 2024) and model bias(Zhu et al., 2024) are developed to achieve this goal. Attention manipulation method would reduce(Huang et al., 2024) or redistribute(Tu et al., 2025) excessive attention scores, so as to steer LVLMs to pay more attention to visual information.

CD methods perform well in hallucination-mitigating tasks but are highly dependent on the quality of the bad samples constructed. If the injured sample still contains a lot of useful information, the contracting operation may cause an even worse result. Attention manipulation methods cost fewer computation resources but are highly sensitive to parameters, presenting unstable performances. Our research constructs bad samples from the perspective of attention, filtering out useful information so that the bad samples only carry the information that needs to be offset, thereby achieving a stable and high-quality effect.

3 Methodology

3.1 Task Formation

Typically, LVLMs aim to generate proper text outputs from multi-modal inputs, especially combined visual and textual data. The visual encoder extracts visual features, then passes them to the modal connection module, where visual features are mapped into the text semantic space. The mapped features are combined with textual tokens, either through concatenation(Liu et al., 2023b) or cross-modal fusion(Dai et al., 2023). The final combined features are then passed into the LLM to generate outputs autoregressively. Formally, given an input image

I, corresponding question text Q, and already generated tokens $y_{<t}$, the next token y_t is decoded according to the probability distribution:

$$p(y_t) = p_\theta(y_t \mid I, Q, y_{<t}) \tag{1}$$

where θ represents the parameters of the LVLM. The goal of hallucination mitigation is to make output sequences contain less contradictory content.

3.2 Formulating image heads masks

LLMs in prevailing LVLMs are most decoder-only, use an attention mechanism to capture the blending of textual and visual features. Formally, the attention matrix A is calculated by:

$$A = \text{softmax}(\frac{Q \cdot K^T}{\sqrt{d_k}}) \tag{2}$$

where Q and K denotes queries and keys respectively, d_k represents the dimension of key vectors. Each row of the attention matrix A indicates the proportion of attention that the current token has invested in the previously generated token. We believe that the higher the sum of the values obtained by the image tokens in the attention matrix is, the more attention the visual information will receive.

There are multiple attention heads in each layer of the LLM, each of which calculates its own attention matrix. We randomly selected 500 images from the validation set of COCO 2014(Lin et al., 2014), input them into LVLMs with the text 'Please describe this image in detail', then record the sum of the attention scores for the image token in the attention matrix of each head in each layer of the model when each token is generated. With a threshold τ, we obtain the attention head matrix where each element represents how many times this attention head has paid over-threshold attention proportion to image tokens (as shown in Figure 2). After normalization, the non-zero attention heads in the attention head matrix are named image heads. Lastly, by masking the selected image heads, the image head mask is constructed.

Apparently, the number of image heads varies with the change of τ. Table 1 shows the number of image heads of LLaVA-1.5-7b and Qwen-VL-7b(Bai et al., 2023b) given different threshold τ values. Intuitively, if the threshold is too high and too few bad samples are masked, then useful information will still be contained in the bad samples; If the threshold is too low, causing the heads that do not pay much attention to the image to be masked as well, the reduction effect of the CD method on semantic information will be weakened. Therefore, choosing the appropriate threshold is an important issue.

Model	τ	# image heads	proportion
LLaVA-1.5-7b	0.95	192	18.75%
	0.9	238	23.24%
	0.8	315	30.76%
	0.7	364	35.55%
	0.6	424	41.41%
	0.5	506	49.41%
Qwen-VL-7b	0.99	248	24.22%
	0.975	317	30.96%
	0.95	395	38.57%
	0.9	473	46.19%

Table 1: **The number and proportion of image heads corresponding to the variation of τ.** τ represents the threshold of considering "high" attention scores paid on image tokens of a head.

3.3 MaskCD

When using an image head mask to construct a bad sample, the masked heads' attention output will be set to zero. Since this method is equivalent to setting the parameters of the corresponding head to zero, we use θ_m to represent the model where the image heads are masked. However, in actual operation, only the attention value is changed; no model parameter will be modified.

Then MaskCD is formulated as equation 3:

$$p(y_t) = \text{softmax}\Big((1+\alpha)\cdot \text{logits}_\theta(y_t|I,Q,y_{<t}) - \alpha \cdot \text{logits}_{\theta_m}(y_t|I,Q,y_{<t})\Big) \tag{3}$$

where $logits$ represents the value of $p(y_t|I,Q,y_{<t})$ before softmax operation. α is a hyperparameter that controls the intensity of contrast.

By subtracting the output logits of bad samples from the original one, MaskCD enables the final output logits to utilize only the truly useful visual and textual information as much as possible, thereby alleviating the hallucination phenomenon of LVLMs.

4 Experiment Settings

4.1 Benchmarks

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 is used to measure the hallucination proportion of the model's generated texts. It evaluates hallucination on two aspects: CHAIR$_S$ and CHAIR$_I$. The former calculates the proportion of sentences containing hallucinations at the sentence level, while the latter computes the hallucinated ratio at the object level. The two metrics can be formulated as:

$$\begin{aligned} \text{CHAIR}_S &= \frac{|\{\text{sentences w/ hallucinated objects}\}|}{|\{\text{all captions sentences}\}|} \\ \text{CHAIR}_I &= \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all mentioned objects}\}|} \end{aligned} \tag{4}$$

We randomly selected 500 images from the validation set of COCO 2014(Lin et al., 2014) and used the prompt "Please describe this image in detail." to obtain the generated captions.

POPE The Polling-based Object Probing Evaluation (POPE) (Li et al., 2023b) is a benchmark for assessing object hallucination. LVLMs are required to answer formatted questions like "Is there a <object> in the image?" with "Yes" or "No". The answers' yes-no ratio is designed to be 50% 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.

We choose MSCOCO dateset for POPE evaluation. The key evaluation metrics are: Accuracy, Precision, Recall, and F1 score.

MME The Multimodal Large Language Model Evaluation (MME) (Fu et al., 2023) assesses LVLMs using set of comprehensive metrics. MME benchmark contains 14 subsets, so as to evaluate LVLMs' general capabilities. Following the methodologies of (Yin et al., 2023), when presenting the test results of all subsets of MME, we divide them into two groups: hallucination and non-hallucination. The hallucination group includes "existence", "counts", "color" and "position", which evaluate LVLMs at the object and attribute level,s respectively.

4.2 Models

LVLM Models We select LLaVA-1.5-7b and Qwen-VL-7b for evaluation. Each model utilizes Vision Transformer (ViT) as the backbone of its visual encoder, but employs different modal connection module and LLMs. LLaVA-1.5-7b directly projects visual embeddings into semantic space through multi-layer perception(MLP), while Qwen-VL-7b utilizes a position-aware vision-language adapter to compress image features. As for the LLM part, LLaVA-1.5-7b utilizes vicuna as LLM backbone, while Qwen-VL-7b's counterpart is Qwen(Bai et al., 2023a). The LLM backbones are both constructed by 32 layers of decoder blocks, and each layer contains 32 heads, resulting in 1024 heads in total.

4.3 Baseline Methods

We compare MaskCD with three classic and effective hallucination mitigating methods: **VCD**(Leng et al., 2024) uses random Gaussian noise to contaminate the original image, reducing the valid information it contains and thus serving as a bad sample. **M3ID**(Favero et al., 2024) deletes the image for the bad sample input, and slightly changes the contrastive decode function. Both of the above methods belong to the CD category. **OPERA**(Huang et al., 2024) takes advantage of the attention sink phenomenon, punishes overly concentrated attention, and combines it with a retrospection-allocation strategy. It is an attention manipulation method based on a beam search strategy. MaskCD is a CD method that only masks the model's inner values to construct bad samples, distinguishing it from other CD-class methods.

5 Result and Analysis

5.1 Overall Result

CHAIR Table 2 shows the overall results for CHAIR evaluation. MaskCD gained evidently better performance compared with other methods. Specifically, MaskCD lowers CHAIR_s and CHAIR_i by 19.12% and 29.87% for LLaVA-1.5-7b, and achieves 75.59% and 50.57% decrease for Qwen-VL-7b. This indicates that our proposed image heads masks are quite effective in hallucination mitigating. Moreover, MaskCD outperforms VCD and M3ID, demonstrating that the bad samples constructed by masking the image head contain less effective information, thereby achieving better results among similar CD methods.

POPE Table 3 shows the evaluation results of POPE benchmark. For both LLaVA-1.5-7b and Qwen-VL-7b, MaskCD represents comparable performance with OPERA, outperforms the baseline

Method	LLaVA-1.5-7b				Qwen-VL-7b			
	CHAIR_s ↓	CHAIR_i ↓	Precision	F1	CHAIR_s ↓	CHAIR_i ↓	Precision	F1
Baseline	50.20	15.40	72.10	73.50	50.8	17.4	68.3	63.0
VCD	55.6	16.4	71.7	75.2	48.4	16.7	68.1	64.6
M3ID	55.4	15.5	70.6	75.7	39.8	8.8	77.2	75.6
OPERA	45.8	13.5	76.6	77.8	42.3	11.8	75.5	**76.3**
MaskCD	**40.6**	**10.8**	**79.1**	**78.2**	**12.4**	**8.6**	**88.5**	64.3

Table 2: **Results on benchmark CHAIR.** CHIAR_s and CHAIR_i are hallucination ratio evaluation metrics, lower scores represent better performances. The Baseline method denotes the standard decoding. The best performances within each setting are bolded. Comparable but not the best performances are underlined.

Setup	**Method**	**LLaVA-1.5-7b**			**Qwen-vl-7b**		
		Acc. ↑	Recall ↑	F1 ↑	Acc. ↑	Recall ↑	F1 ↑
random	Baseline	82.90	72.07	80.82	81.97	77.67	81.16
	VCD	85.57	76.27	84.09	76.20	**81.73**	77.45
	M3ID	85.27	74.67	85.52	74.60	69.67	73.28
	OPERA	**89.30**	**89.00**	**89.27**	66.33	**81.73**	77.45
	MaskCD	88.77	87.47	88.62	**87.77**	79.47	**86.66**
popular	Baseline	81.10	74.27	79.22	80.20	78.40	79.84
	VCD	83.67	72.34	82.36	72.30	**81.80**	74.70
	M3ID	83.60	73.77	81.99	72.07	70.33	71.57
	OPERA	**85.93**	**87.96**	**86.86**	66.77	73.67	68.91
	MaskCD	85.67	87.53	85.83	**86.57**	79.40	**85.53**
adversarial	Baseline	78.60	72.35	77.10	78.43	78.60	78.47
	VCD	81.07	76.24	80.11	71.57	**83.07**	71.50
	M3ID	**81.57**	73.36	80.20	71.83	70.67	71.50
	OPERA	79.00	**88.03**	80.91	67.50	73.67	69.38
	MaskCD	79.63	87.53	**81.12**	**83.40**	79.27	**82.68**
All	Baseline	80.87	72.90	79.05	80.20	78.22	79.82
	VCD	83.44	74.95	82.19	73.36	**82.20**	75.55
	M3ID	83.48	73.93	82.57	72.83	70.22	72.12
	OPERA	84.74	**88.33**	**85.68**	66.87	73.67	68.97
	MaskCD	**84.88**	88.20	85.48	**85.91**	79.38	**84.96**

Table 3: **Results on benchmark POPE.** The Baseline method denotes the standard decoding. The best performances within each setting are bolded. Comparable but not the best performances are underlined.

Method	Objectlevel		Attributelevel		Total
	existence	count	position	color	
Baseline	180.00	101.67	100.00	153.33	535.00
VCD	175.00	106.67	111.67	146.67	540.01
M3ID	180.00	101.67	105.00	**158.33**	545.00
OPERA	**195.00**	148.33	128.33	155.00	626.66
MaskCD	**195.00**	**168.33**	**133.33**	150.00	**646.66**

Table 4: **Results on benchmark MME (four hallucination subsets) of LLaVA-1.5-7b.** The Baseline method denotes the standard decoding. The best performances within each setting are bolded. Comparable but not the best performances are underlined.

Method	Object-level		Attribute-level		Total
	existence	count	position	color	
Baseline	105.00	83.33	50.00	136.67	375.00
VCD	86.67	91.67	41.67	111.67	346.00
M3ID	**118.33**	**98.33**	51.67	125.00	331.68
OPERA	93.67	84.33	46.67	121.33	393.33
MaskCD	111.67	95.00	**65.00**	**138.33**	**410**

Table 5: **Results on benchmark MME (four hallucination subsets) of Qwen-VL-7b.** The Baseline method denotes the standard decoding. The best performances within each setting are bolded. Comparable but not the best performances are underlined.

and other CD methods, indicating its excellence in hallucination alleviation. Furthermore, as taking computational cost into account, MaskCD achieves a similar performance effect with a computational

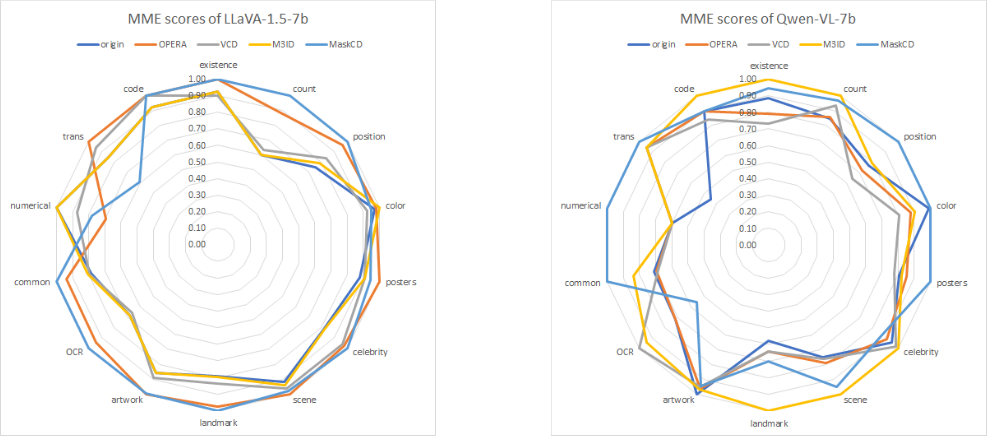

Figure 3: **Visualization of MME scores of LLava-1.5-7b(left) and Qwen-VL-7b(right).** Scores are normalized by dividing maximum score of each subset.

cost lower than that of OPERA, demonstrating its excellence.

MME Table 4 and Table 5 shows the results on four hallucination-related MME subsets for LLaVA-1.5-7b and Qwen-VL-7b, respectively. MaskCD achieves best performances on every single subset for LLaVA-1.5-7b and on attribute-level subsets for Qwen-VL-7b. The evaluation on object-level subsets of Qwen-VL-7b also achieves the second-best results, representing the effectiveness of MaskCD in alleviating hallucinations. Meanwhile, Figure 3 shows the overall results for all 14 subsets of the MME benchmark. It is evident that besides the capability of mitigating hallucination, MaskCD also retains or even partially improves the model's ability in general evaluation.

5.2 Ablation Study

In this subsection, we present ablation studies to examine the impact of mask selecting and other hyper-parameters. We conduct these experiments with LLaVA-1.5-7b.

Mask selection To demonstrate the necessity of masking the image heads rather than other heads, for each image head mask in the settings, we randomly select an equal number from other heads to form a random mask. The method of using these random masks for MaskCD is denoted as MaskCD_r. Table 6, 7, and 8 show the performance of MaskCD and MaskCD_r on CHAIR, POPE and MME, respectively. The results show that masking random heads also has a slight effect on alleviating hallucinations, but it cannot compete with the results of masking image heads. It indicates that the image heads indeed contain more useful and necessary information, so it is rational to mask them rather than other heads.

Method	CHAIR_s ↓	CHAIR_i ↓	Pre.	F1
Baseline	50.2	15.4	72.1	73.5
MaskCD	**40.6**	**10.8**	**79.1**	**78.2**
MaskCD_r	44.0	14.3	77.3	74.0

Table 6: **Results of MaskCD and MaskCD_r on CHAIR evaluation.** "Pre." is the abbreviation for "Precision". It is evident that MaskCD_r indeed helps mitigate hallucination, but cannot compete with MaskCD.

	Method	Accuracy	Precision	Recall	F1
All	Baseline	80.87	87.62	72.07	79.05
	MaskCD	**84.88**	83.11	**88.20**	**85.48**
	MaskCD_r	83.61	**90.67**	75.53	82.38

Table 7: **Results of MaskCD and MaskCD_r on POPE benchmark.** MaskCD_r performs best on Precision metrics, but fails in Recall.

Method	Object-level		Attribute-level	
	existence	count	position	color
Baseline	180.00	101.67	100.00	**153.33**
MaskCD	**195.00**	**168.33**	**133.33**	150.00
MaskCD_r	190.00	133.33	123.33	150.00

Table 8: **Results of MaskCD and MaskCD_r on the hallucination-related subsets of MME.** MaskCD_r indeed helps slightly in mitigating hallucinations.

Mask proportion and CD tensity As mentioned in section 3.2 and 3.3, there are two important hyper-parameters in MaskCD: τ, as the threshold for determining the image head, controls the number of masked heads; and α, which controls the intensity of contrastive decoding operation.

We conduct ablation experiments of τ and α on

Method	τ	CHAIR_s	CHAIR_i	F1
Baseline	/	50.2	15.4	73.5
MaskCD	0.95	46.8	12.9	77.0
	0.9	**40.6**	**10.8**	**78.2**
	0.8	40.8	14.7	74.8
	0.7	48.4	13.0	76.4
	0.6	49.6	14.0	75.9
	0.5	54.8	14.4	75.2

Table 9: **Results of MaskCD with different threshold τ.** It can be seen that the value of τ that is either too small or too large is not conducive to dealing with hallucination problems.

Method	α	CHAIR_s	CHAIR_i	F1
Baseline	/	50.2	15.4	73.5
MaskCD	0.5	43.6	11.5	**78.4**
	1.0	**40.6**	**10.8**	78.2
	2.0	45.2	12.2	77.3
	3.0	44.6	11.6	77.3
	4.0	42.2	11.5	77.6
	5.0	41.6	11.7	77.9
	6.0	41.2	11.8	77.9

Table 10: **Results of MaskCD with different α.** It can be seen that even when α takes a large value, MaskCD can still operate stably, effectively alleviating the hallucination phenomenon.

LLaVA-1.5-7b with CHAIR evaluation. Table 9 shows the results of MaskCD with different thresholds τ. It indicates that the best value of τ is 0.9, which means around 23% of the heads in LLaVA-1.5-7b's LLM backbone are recognized as image heads and have been masked (according to Table 1). Whether too small or too big the value of τ is, the performances tend to decline, and even fail the baseline when τ is 0.5. This shows that the heads to be masked should be delicately selected, and MaskCD achieves this successfully.

Meanwhile, Table 10 shows the results of MaskCD on CHAIR with different α. α is a common hyperparameter in contrastive decoding approaches, whose value controls the intensity of the contrast operation. It can be seen that even when the value of α is quite large, MaskCD can still operate stably and effectively alleviate the hallucination phenomenon. It demonstrates the stability, reliability and practicability of MaskCD as a CD method.

6 Conclusion

In this paper, we first introduce the image heads: the heads in LVLM's LLM backbone that tend to pay comparably high attention proportion on image tokens. Then we propose the image head Masked Contrastive Decoding (MaskCD) method, a novel contrastive decoding approach featuring in masking image heads to construct contrastive samples. MaskCD constructs the contrastive samples of CD methods from the perspective of attention score, combining the effectiveness and stability of these two methods. Extensive experimental results on CHAIR, POPE and MME demonstrate the effectiveness and stability of MaskCD in mitigating the phenomenon of hallucinations. We hope this work can provide a new perspective for exploring future efforts to alleviate hallucinations in LVLMs.

Limitations

Although MaskCD achieves significant performance in hallucination mitigation, it still has several limitations. First, MaskCD requires the use of images for inference in advance to obtain the masks of image heads, which occupy computing resources. Secondly, although the process of obtaining the mask is simple, the obtained mask is only applicable to the same family of LLM backbones. For new LLM bases, the corresponding masks need to be re-obtained. This limitation encourages us to explore how to dynamically construct masks during the model's operation, so as to get rid of these restrictions.

Ethical Considerations

The main research objects of this work are alleviating hallucination phenomenon, which help avoid disloyal contents generated by LVLMs. Moreover, we conduct experiments on the public datasets, which do not contain any offensive content or information with negative social impact. Our research contents are completely in line with the ethical review.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian

Borgeaud, and 8 others. 2022. Flamingo: a visual language model for few-shot learning. In *NeurIPS*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023a. Qwen technical report. *CoRR*, abs/2309.16609.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-vl: A frontier large vision-language model with versatile abilities. *CoRR*, abs/2308.12966.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. 2024. DRESS : Instructing large vision-language models to align and interact with humans via natural language feedback. In *CVPR*, pages 14239–14250. IEEE.

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

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*.

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 *CVPR*, pages 14303–14312. IEEE.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. MME: A comprehensive evaluation benchmark for multimodal large language models. *CoRR*, abs/2306.13394.

Hongyu Hu, Jiyuan Zhang, Minyi Zhao, and Zhenbang Sun. 2023. CIEM: contrastive instruction evaluation method for better instruction tuning. *CoRR*, abs/2309.02301.

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 *CVPR*, pages 13418–13427. IEEE.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *EMNLP*, pages 9161–9175. Association for Computational Linguistics.

Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. 2024. Volcano: Mitigating multimodal hallucination through self-feedback guided revision. In *NAACL-HLT*, pages 391–404. Association for Computational Linguistics.

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 *CVPR*, pages 13872–13882. IEEE.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023a. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *EMNLP*, pages 292–305. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *ECCV (5)*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023a. Aligning large multi-modal model with robust instruction tuning. *CoRR*, abs/2306.14565.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024a. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *ICLR*. OpenReview.net.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023b. Improved baselines with visual instruction tuning. *CoRR*, abs/2310.03744.

Shi Liu, Kecheng Zheng, and Wei Chen. 2024b. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *ECCV (83)*, volume 15141 of *Lecture Notes in Computer Science*, pages 125–140. Springer.

Jiahao Nie, Gongjie Zhang, Wenbin An, Yap-Peng Tan, Alex C. Kot, and Shijian Lu. 2024. Mmrel: A relation understanding dataset and benchmark in the MLLM era. *CoRR*, abs/2406.09121.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

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 *ICML*, 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 *EMNLP*, pages 4035–4045. Association for Computational Linguistics.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2024. Aligning large multimodal models with factually augmented RLHF. In *ACL (Findings)*, pages 13088–13110. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Chongjun Tu, Peng Ye, Dongzhan Zhou, Lei Bai, Gang Yu, Tao Chen, and Wanli Ouyang. 2025. Attention reallocation: Towards zero-cost and controllable hallucination mitigation of mllms. *CoRR*, abs/2503.08342.

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

Sangmin Woo, Jaehyuk Jang, Donguk Kim, Yubin Choi, and Changick Kim. 2024. RITUAL: random image transformations as a universal anti-hallucination lever in lvlms. *CoRR*, abs/2405.17821.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. In *ICLR*. OpenReview.net.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *CoRR*, abs/2311.04257.

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. *CoRR*, abs/2310.16045.

Yifan Zhang, Weichen Yu, Qingsong Wen, Xue Wang, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. 2024. Debiasing multimodal large language models. *CoRR*, abs/2403.05262.

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

Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. 2024. IBD: alleviating hallucinations in large vision-language models via image-biased decoding. *CoRR*, abs/2402.18476.