Octopus: Alleviating Hallucination via Dynamic Contrastive Decoding

Wei Suo[1,2,4*], Lijun Zhang[1,2,4*], Mengyang Sun[2,3,4], Lin Yuanbo Wu[5], Peng Wang[1,2,4†], Yanning Zhang[1,2,4]
[1]School of Computer Science, [2]Ningbo Institute, [3]School of Cybersecurity,
Northwestern Polytechnical University, China.
[4]National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean.
[5]Department of Computer Science, Swansea University, United Kingdom.
{suowei1994,lijunzhang,sunmenmian}@mail.nwpu.edu.cn {peng.wang,ynzhang}@nwpu.edu.cn

Abstract

Large Vision-Language Models (LVLMs) have obtained impressive performance in visual content understanding and multi-modal reasoning. Unfortunately, these large models suffer from serious hallucination problems and tend to generate fabricated responses. Recently, several Contrastive Decoding (CD) strategies have been proposed to alleviate hallucination by introducing disturbed inputs. Although great progress has been made, these CD strategies mostly apply a one-size-fits-all approach for all input conditions. In this paper, we revisit this process through extensive experiments. Related results show that hallucination causes are hybrid and each generative step faces a unique hallucination challenge. Leveraging these meaningful insights, we introduce a simple yet effective Octopus-like framework that enables the model to adaptively identify hallucination types and create a dynamic CD workflow. Our Octopus framework not only outperforms existing methods across four benchmarks but also demonstrates excellent deployability and expansibility. Code is available at https://github.com/LijunZhang01/Octopus.

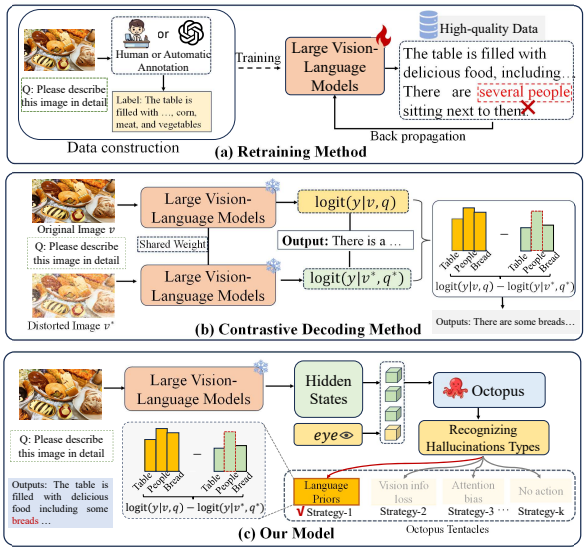

Figure 1. Paradigm comparison of different hallucination alleviation methods. (a) Retraining method. Constructing high-quality data to retrain these LVLMs. (b) Contrastive Decoding. Comparing the output distributions from the original and distorted inputs. (c) Octopus. Our method focuses on dynamically selecting suitable strategies to reduce hallucinations caused by various factors.

1. Introduction

Large Vision-Language Models (LVLMs) have achieved significant success over the past few years [4, 7, 25, 26, 61]. They have facilitated various Vision-and-Language (VL) tasks [14, 22, 40, 41] by adapting to different input instructions. However, LVLMs are facing a grand challenge: they often fail to accurately capture the visual content and tend to generate fabricated responses (*e.g.,* imaginary objects, incorrect attributes and inexistent relationship), which is known as *hallucination* [13, 24]. The hallucination issue seriously affects user trust and confidence, especially in applications that require trustworthy outcomes such as medical reports [10, 52] and automatic driving [8].

To alleviate the hallucination issue, existing approaches can be roughly categorized into two research lines. As shown in Fig. 1 (a), the first paradigm relies on constructing high-quality instruction tuning data and re-training the models to suppress hallucinations [13, 24, 28, 44]. However, such a strategy requires a well-designed data construction process with complex verification and expensive costs. In addition, additional training is strictly prohibited for deployed models.

In contrast, as shown in Fig. 1 (b), Contrastive Decod-

*These authors contributed equally to this work.
†Corresponding authors.

ing (CD) methods [11, 19, 29, 51] aim to contrast the logit scores at each generative step from the original inputs (*i.e.,* image input v and corresponding text q) with those derived from the modified inputs (*i.e.,* v^* and q^*). As a post-hoc corrective mechanism, such a methodology can effectively eliminate hallucination without complex training. In practice, these CD methods focus on designing a new pair of v^* and q^* to alleviate hallucination problems. For example, [19] utilizes Gaussian noise to mask the original image and overcome language priors. [51] employs enhanced visual input to mitigate attention bias.

Despite their strong performance, these CD methods mostly apply the same disturbed manner for all samples and generative steps. This motivates us to question: 1) Is a single CD strategy suitable for different samples? 2) Do all generative steps (*i.e.,* predicted tokens) experience the same type of hallucination? To answer the two questions, we conduct exploratory experiments to understand the causes of hallucinations at both sample and token levels. Specifically, we use a group of CD methods as off-the-shelf diagnostic tools to investigate the pattern of hallucinations. Our analysis reveals that each CD method is only effective on specific hallucinated samples, and using a single CD strategy would lead to sub-optimal results. Meanwhile, we thoroughly investigate the process of hallucination emergence at the token level. Through an enumeration method [5] and qualitative analysis, we find that fabricated responses are hybrid and each generative step faces a unique hallucination issue.

The above results indicate that a one-size-fits-all approach struggles to correct different types of hallucinations effectively. Thus a natural idea is to combine multiple CD methods to reduce hallucinations from various sources. However, without well-defined labels, identifying the optimal strategy for different input samples is challenging. Additionally, token generation is sequentially dependent and involves a vast solution space, making it difficult to choose the best CD approach at each generative step.

To tackle the above problems, as shown in Fig.1 (c), we introduce a simple yet effective framework, called **Octopus**. Different from previous works, our method focuses on guiding the model to dynamically organize the contrastive decoding workflow and selecting the appropriate CD strategy based on different inputs. In particular, we first build a transformer-based block and a learnable token to adaptively recognize the type of hallucination, similar to the Octopus's eyes. According to different decisions, each CD strategy is regarded as a "tentacle" to perform a specific contrastive operation. Finally, leveraging Direct Preference Optimization (DPO) [34] or reinforcement learning [37, 49], Octopus can be easily optimized. Benefiting from the above designs, the proposed method not only effectively reduces hallucinated content, but also scales well for deployments due to avoiding retraining weights of LVLMs. More importantly, as a general framework, subsequent CD methods can be seamlessly integrated without additional adjustments. In summary, we make the following contributions:

1) Our work reveals that the mechanism of hallucination occurrence is a complex hybrid and different samples (or tokens) suffer from various forms of hallucination challenges.

2) We develop a new framework Octopus that can adaptively recognize the types of hallucination and build a dynamically contrastive decoding workflow to correct fabricated content.

3) Octopus achieves state-of-the-art performance across four benchmarks for both generative and discriminative tasks, while also demonstrating excellent deployability and expansibility.

2. Related work

2.1. Large Visual-Language Models

Large Visual-Language Models (LVLMs) [4, 7, 25, 26, 61] usually consist of three key components: a visual encoder like CLIP [33], a Large Language Model (LLM) such as LLAMA [42] and a cross-modal alignment module that connects the visual encoder's output to the LLM. LVLMs have obtained impressive performance in visual content understanding and multi-modal reasoning such as image captioning [14, 61], referring expression comprehension [27, 56], human-object interaction [58], and visual question answering [9, 46].

2.2. Hallucination in LVLMs

To alleviate the hallucinations in LVLMs, both data-driven retraining and Contrastive Decoding (CD) methods have been proposed. Data-driven methods aim to enhance data quality to reduce the hallucinations [13, 24, 28, 44]. For example, some works introduce negative data [24] and counterfactual data [54] to mitigate hallucination issues. [48] cleans the dataset to minimize noise and errors. [50, 55] annotates a high-quality hallucination dataset to suppress the occurrence of hallucinations by fine-tuning. In contrast, CD methods tackle hallucinations by comparing output distributions from original and distorted input without altering the model's weights. For instance, [19] alleviates hallucinations by counteracting language priors, while [11] tackles them through refined visual prompts. Different from the above methods, our work focuses on adaptively selecting the most suitable CD strategies to alleviate different hallucination issues.

3. Preliminary

3.1. Large Vision Language Models

Given a Large Vision-Language Model (LVLM) with parameters θ, the model can effectively perform various multi-modal tasks using a visual input v and a textual instruction

q. Specifically, at each generative step t, the auto-regressive LVLM conducts the following calculations:

$$\ell_t = \log p(y_t|v, q, y_{<t}; \theta), \tag{1}$$

$$y_t \sim \text{Softmax}(\ell_t), \tag{2}$$

where ℓ_t is the logit score for the next token y_t, while $y_{<t}$ represents the response generated before time step t. After extensive training, well-designed LVLMs demonstrate impressive understanding ability across a wide range of multi-modal tasks. However, these models suffer from serious hallucination issues [13, 24, 28, 44]. They often produce inaccurate answers or fabricated descriptions that may not align with the visual input.

3.2. Contrastive Decoding

To mitigate cross-modal hallucinations, Contrastive Decoding (CD) offers a promising approach by contrasting output distributions between original and distorted inputs. In particular, CD methods first generate two output distributions: one from the original visual image v and textual query q, and another from the perturbed inputs v^* and q^*. Then, by examining the difference between two distributions, a contrastive response ℓ_{cd} can be formulated as follows:

$$\begin{aligned}\ell_{cd} = m&\log p(y_t|v, q, y_{<t}; \theta) - \\ n&\log p(y_t|v^*, q^*, y_{<t}; \theta),\end{aligned} \tag{3}$$

$$y_t \sim \text{Softmax}(\ell_{\text{cd}}), \tag{4}$$

where m and n are hyperparameters, and y_t represents the predicted token based on the contrastive decoding. Although CD methods are effective in mitigating hallucinations, they generally apply the same disturbed manner to all samples and generative steps. In this paper, we rethink the above operations with extensive experiments. Considering that existing CD methods are often tailored to specific types of hallucinations, we select three well-received CD methods as diagnostic tools to further investigate the mechanisms underlying hallucination occurrence. We will introduce them in detail.

Strategy-1: VCD [19]. VCD focuses on ***overcoming language priors***. It employs a Gaussian noise mask to generate the distorted visual input v^*, while the query text q is unchanged.

Strategy-2: M3ID [11]. M3ID relieves hallucinations by ***reducing visual information loss***. It masks the query text to build q^* and independently supplies the visual input v into the LVLMs as the distorted input.

Strategy-3: AVISC [51]. AVISC alleviates hallucinations by ***minimize attention bias***. It constructs a new visual input v^* using blind tokens that do not contain information relevant to the query.

These three strategies correspond to different causes of hallucinations: ***language priors, vision information loss and attention bias***. Next, we conduct related experiments on the popular VL model LLaVA-1.5-7B [25] to explore the two key questions: 1) Is a single CD strategy effective for all multi-modal samples? 2) Does each time step experience the same cause of hallucination?

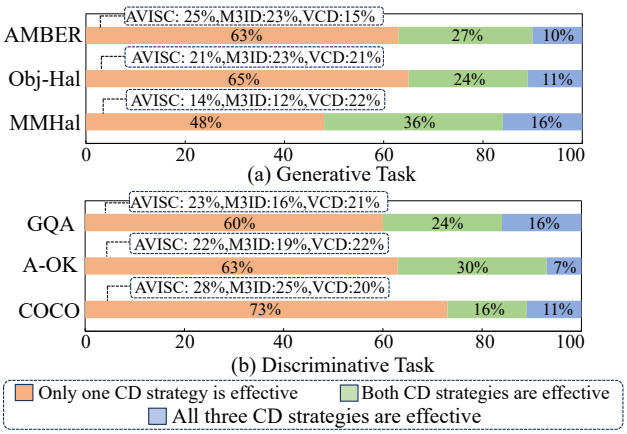

Figure 2. The proportion of effective samples using different CD methods for (a) Generative Task and (b) Discriminative Task. We observe that each CD strategy can only address part of the samples.

3.3. Sample-Level Hallucination

In this section, we conduct two kinds of experiments (*i.e.,* generative and discriminative tasks) to answer the first question. For generative task, we establish experiments on the three widely used datasets: AMBER [47], Object-HalBench [35] with the language prompt "Please describe this image in detail", as well as MMHalbench [39] with the original instructions as the prompts. Following [45], both the AMBER and Object-HalBench use CHAIR score [35] as metrics to evaluate the degree of hallucination, while the MMHalbench uses the GPT-4V score [39]. In addition, we use these three strategies in Sec. 3.2 to interfere with the output distributions of LLaVA for each sample, respectively. By utilizing the above metrics, this strategy is identified as effective when it attains better performance compared to the original LLaVA output. As shown in Fig. 2 (a), we report the corresponding percentages, where the orange, green and blue denote "Only One CD strategy is effective", "Both CD strategies are effective" and "All three CD strategies are effective", respectively. By comparing these results, we observe that a large number of samples (∼60%) can only be addressed by certain specific CD strategies and their overlap is relatively small (∼10%).

For discriminative task, we conduct similar experiments on the three POPE datasets [21] (*i.e.,* GQA [16], A-OKVQA [38] and COCO [23]) with the language template "Q + Please answer this question in one word", where Q represents the textual question. Different from the generative task, we directly apply strategies 1-3 to each question

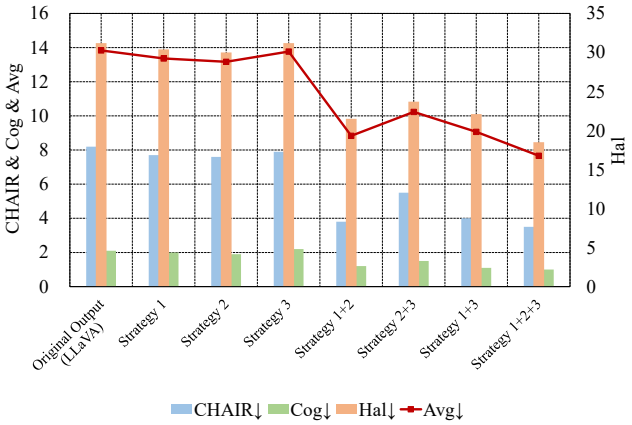

Figure 3. Token-level hallucination quantitative evaluation. We enumerate different CD strategies at each time step. The results show that using multiple CD strategies obtains better performance.

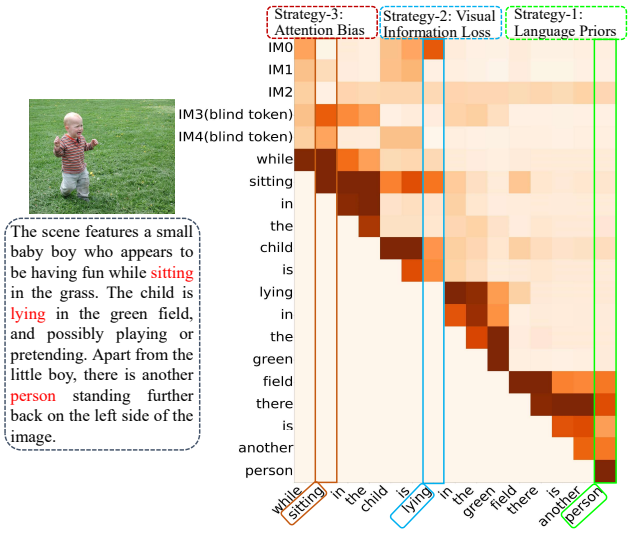

Figure 4. Token-level hallucination qualitative analysis. For simplicity, we only present the attention map across the top 5 visual tokens and corresponding keywords. The results show that hallucination causes are hybrid in a sample.

and count the number of samples whose answers are corrected. As shown in Fig. 2 (b), we report the corresponding percentages and observe that each CD strategy addresses a subset of the samples, and only ∼10% of cases are effective across all three methods. Based on the above results, we conclude that ***each CD method is only effective on specific hallucinated samples, and using a single strategy for all cases would inevitably lead to sub-optimal results.***

3.4. Token-Level Hallucination

The above sample-level analyses demonstrate that each sample needs to adopt a specific CD strategy. In this section, we focus on a more fine-grained scenario: whether each time step in the generative process suffers from the same hallucination causes.

To counteract this, we construct experiments on the AMBER [47] dataset with three metrics: CHAIR [35], Cog [47] and Hal [47]. Because these CD strategies in Sec. 3.2 can be regarded as three types of diagnostic tools, we apply the enumeration method [5] to evaluate the hallucinatory causes in the generative process. Moreover, we are aware that even though there are just three CD candidates, the combination space is still enormous due to the lengthy outputs. To reduce the number of combinations, the enumerating space would only consider the first three hallucinated nouns in each description. As shown in Fig.3, we use "strategy-1", "strategy-2" and "strategy-3" to denote the hallucination mitigation strategies introduced by Sec. 3.2 (*i.e.,* VCD [19], M3ID [11] and AVISC [51]). Meanwhile, we exhibit the best scores from these combinations. Take "strategy 1+3" as an example, each of the three tokens has two selectable hallucination elimination strategies (*i.e.,* strategy-1 and strategy-3), thus there are a total of 6 combinations. For simplicity, we only report the best results among these combinations. By comparing these scores, we find that leveraging multiple CD strategies can better suppress hallucinations.

To investigate the nature of hallucination occurrence, we also conduct a qualitative analysis to examin the attention distribution for each predicted token. As shown in Fig. 4, it can be found that hallucinated words include "sitting, lying, and person" in this sentence, where each token corresponds to different causes of hallucination. For example, the "sitting" focused on the visual blind token "IM3", indicating that the current step is affected by attention bias [51]. The occurrence of "lying" is primarily due to insufficient attention to visual information [11]. While the "person" concentrates solely on language tokens, suggesting that it is influenced by language priors [19]. Therefore, we conclude that ***the hallucination causes are hybrid and each generative step faces different forms of challenge.***

3.5. Discussion

Based on the above experiments, it can be found that various hallucination factors collectively lead to falsity outputs across the sample and token levels. Therefore, a natural idea is to combine these off-the-shelf CD strategies as corrective approaches and tackle different types of hallucinations. However, due to lacking pre-defined labels, it is difficult to select the most suitable strategy for each sample. Meanwhile, since the textual generation process is sequentially dependent, the choice of hallucination elimination strategy for each token would be influenced by the previous selection. Considering the vast solution space, deciding which CD strategy to use is challenging at each generative step.

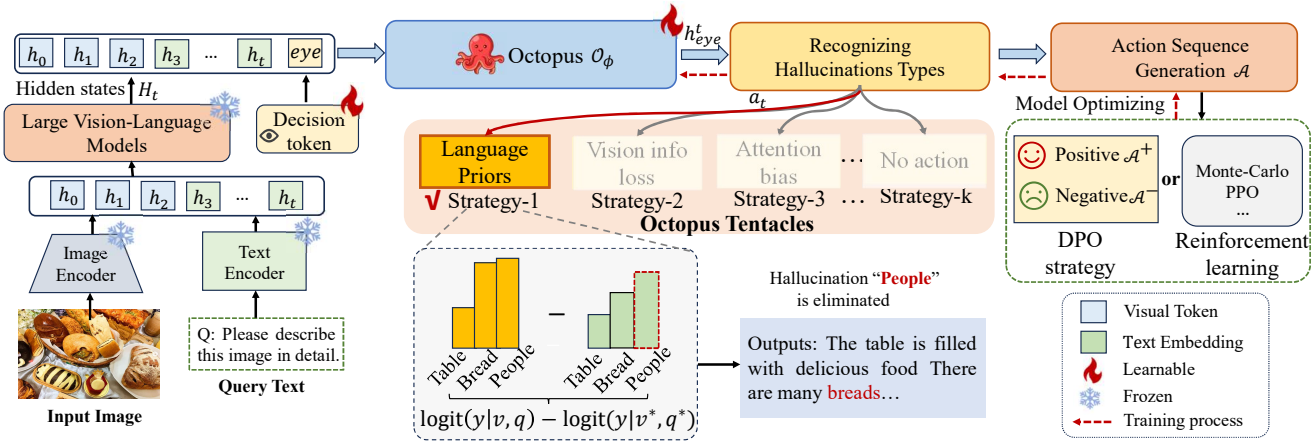

Figure 5. **Overview of our method.** Our Octopus framework consists of two key components: the decision token eye and its tentacles. Specifically, we first utilize the "eye" to identify the types of hallucinations, and then these "tentacles" are applied to address specific hallucination issues at each generative step. Finally, our model would be optimized by DPO or other reinforcement learning methods.

4. Our Method

To alleviate the above challenges, we propose a simple yet effective framework Octopus, which can adaptively select the proper CD strategy and organize the contrastive decoding workflow based on different input conditions. As shown in Fig.5, instead of relying on just one technology, our method focuses on leveraging an Octopus-like structure to detect hallucination types and perform corresponding contrastive actions. Considering that the discriminative task can be viewed as a generation sequence with a response length of 1, we will take the more complex generative task as an example to introduce our method.

4.1. Model Structure

Given a vision input v and a textual instruction q (*e.g.,* "Please describe this image in detail"), we utilize the Octopus's eye to recognize the type of hallucination at each generative step. Then these Octopus's tentacles are used to perform specific strategies by contrastive decoding.

Specifically, we first construct a vanilla transformer-based block $\mathcal{O}_\phi$, where ϕ denotes the parameters of the transformer structure [43]. Based on Eq.1, it can be found that each time step y_t would be influenced by v, q, and $y_{<t}$ together. Therefore, these hidden states from LVLMs (*i.e.,* v, q and $y_{<t}$) would be fed into the $\mathcal{O}_\phi$ with a decision token $eye \in \mathbb{R}^d$, where d is the dimension of hidden state. For simplicity, we use $H_t = \{h_i\}_{i=1}^t$ to represent the sequence before t-th generation step, where $h_i \in \mathbb{R}^d$ is hidden state of each token. While the learnable token eye can be regarded as "Octopus's eye". Formally, the above computations can be formulated as:

$$[h^t_{eye}; H^{'}_t] = \mathcal{O}_\phi\ (\text{concat}[eye; H_t] + E_{pos}), \tag{5}$$

where h^t_{eye} and $H^{'}_t$ are corresponding outputs from eye and H_t sequence, respectively. While E_{pos} and concat denote position embedding and concatenate operation. Benefiting from the self-attention mechanism [43], the h^t_{eye} can adaptively aggregate the information from other hidden states.

Then, a light and simple Multi-Layer Perceptron (MLP) is utilized to map the h^t_{eye} into action vector $h^t_{act} \in \mathbb{R}^k$, where the k is the number of candidate strategies. In this paper, we build four action spaces at each step including strategies 1-3 in the Sec. 3.2 (*i.e.,* VCD [19], M3ID [11] and AVISC [51]) and a null action (*i.e.,* no CD strategy is performed). Here, we use "tentacles" to represent these candidate CD actions. For each h^t_{act}, the action vector a_t is obtained by:

$$h^t_{act} = \text{MLP}(h^t_{eye}), \tag{6}$$

$$a_t = \text{argmax}(\text{Softmax}(h^t_{act})), \tag{7}$$

where argmax refers to the operation of selecting the index of the maximum value, while Softmax is the active function. Based on the one-hot vector a_t, our Octopus can conveniently choose the corresponding CD strategies to implement. Finally, we would obtain a contrastive decoding workflow $\mathcal{A} = \{a_t\}_{t=1}^N$, where N is the length of response.

4.2. Model Optimizing

It can be noted that there is a non-differentiable operation in the above computations (*i.e.,* Eq.7), and the optimization process is also challenging due to the lack of explicit decision labels and serious curse of dimensionality [1]. Therefore, we introduce Direct Preference Optimization (DPO) [34] to alleviate this problem. In fact, our Octopus can also be optimized using other Reinforcement Learning (RL) methods such as Monte-Carlo sampling [49] or PPO [37] (more discussion can be found in Sec. 5.4).

Due to its simplicity and stability compared to other approaches [37, 49], we will only introduce the DPO optimization approach in this section.

Data Construction. The DPO method is designed to replace the typical RLHF procedure [31] and it can directly fit the human preference by building positive and negative samples. Inspired by this, we reformulate the above action choice process as a preference problem, which encourages our Octopus $\mathcal{O}_\phi$ to produce the action sequences that can effectively mitigate hallucinations. For constructing the positive workflow $\mathcal{A}^+$ and negative workflow $\mathcal{A}^-$, we generate 10 sequences for each sample by randomly selecting four actions at each generative step. Next, we divide them into $\mathcal{A}^+$ and $\mathcal{A}^-$ from these responses according to the CHAIR metric [35]. In practice, this metric is also flexible and can be adjusted for different fields. For the discrimination task, we separately use four "tentacles" to build $\mathcal{A}$, while the positive and negative samples are split by the answer's confidence scores. Finally, following [20], we use balanced positive and negative samples to construct the preference dataset.

Training Process. To guide the policy model output preferred sequences $\mathcal{A}^+$, the traditional RL-based methods have to depend on a complex sampling process [49] or an additional reward model [37]. Conversely, the DPO replaces the reward process with a policy model and a reference model, which can straightforwardly increase the likelihood of positive sequences. Therefore, given the above preference dataset $\mathcal{D} = \{\mathcal{A}^+, \mathcal{A}^-\}$, we apply the DPO to directly optimize our Octopus. In addition, previous works have proved that removing the reference model can obtain better or comparable performance than the original DPO [30, 53]. Based on this, our optimization objective is defined as follows:

$$\max_{\mathcal{O}_\phi} \mathbb{E}_{(x,\mathcal{A}^+,\mathcal{A}^-)\sim\mathcal{D}} \log \sigma \left(\beta \log \mathcal{O}_\phi \left(\mathcal{A}^+ \mid x\right) - \beta \log \mathcal{O}_\phi \left(\mathcal{A}^- \mid x\right)\right), \tag{8}$$

where $x = (v, q)$ is the input sequence, σ denote sigmoid function. Following [20], we set the β to 1. Based on the above training process, our Octopus can adaptively learn to construct a suitable workflow without human labeling. Moreover, note that our method would only optimize the parameters ϕ of the Octopus, the weights of LVLMs would remain frozen.

5. Experiment

5.1. Experimental Setting

Datasets. We conducted experiments on both generative and discriminative tasks to study the hallucinations of LVLMs. Following previous methods [11, 19, 45, 51], we mainly build the experiments for the generative task on the AMBER [47], Object-HalBench [35], and MMHal-Bench [39] datasets. For the discriminative task, we evaluate the results on the AMBER [47] and POPE [21] datasets.

Evaluation Metric. Following [35, 39, 51], we use four metrics to evaluate generative hallucinations on the AMBER and Object-HalBench including CHAIR [35], Cover [47], Hal [47], and Cog [47]. While for the MMHal-Bench dataset, we use GPT-4 [2] to evaluate the quality of responses. For the discrimination task, we follow [51] and use Accuracy and F1 to measure hallucinations.

Implementation Details. To train our model, we construct two datasets for the generation and discrimination tasks. For the generation task, we build 10,000 preference data on MSCOCO [23] with a language prompt "Please Describe this image in detail.". For the discrimination task, we follow [21] to build 7,000 hallucinated data from the MSCOCO [23] training set. The Adam [18] is used as our optimizer. We train all models on the four 3090 GPUs and the batch size is set to 4.

5.2. Quantitative Evaluation

Generative Task. In Table 1, we show a performance comparison on the generative task, related results from [45, 51]. Two general LVLMs (*i.e.,* LLaVA [25] and InstructBLIP [7]) are applied to evaluate the results across three datasets including AMBER [47], Object-HalBench [35] and MMHalBench [39]. In particular, we observe that our Octopus can significantly boost performance on all datasets compared with previous CD methods [11, 19, 51]. Meanwhile, compared to the original LLaVA model (referred to as "Base"), our approach achieves ∼40% performance improvement on the CHAIR metric of the AMBER dataset. Moreover, we also report the results for approaches that require retraining the entire model [36, 57, 59], it can be found that Octopus still outperforms them by a large margin.

Discriminative Task. To verify the effectiveness of our method on the discriminative task, we conduct experiments on the AMBER and POPE datasets. As shown in Tabel 2, our model achieves significant performance gains and boosts the baseline model 9.7/11.6 and 3.75/3.02 in accuracy and F1 score for two benchmarks, respectively.

Finally, note that the purpose of our work is not to surpass all methods across every benchmark, and we believe there remains significant room for improvement in the future by integrating more effective CD strategies.

5.3. Ablation Study

As shown in Table 3, we conduct several ablation studies on the AMBER to demonstrate the effectiveness of our method. In the first two rows, we report the scores of the original LLaVA and corresponding results based on randomly using three CD strategies at each generative step. The results show that the randomly selected CD strategy

Table 1. Comparison with the state-of-the-art methods for the generative task across three datasets. For reference, we also provide the results of fine-tuning LVLMs. † signifies results reproduced with the official implementation codes.

LVLMs	Method	MMHalBench		Object HalBench		AMBER			
		Score ↑	HalRate ↓	$CHAIR_s$ ↓	$CHAIR_i$ ↓	CHAIR ↓	Cover. ↑	HalRate ↓	Cog. ↓
Referenced Results (Not Directly Comparable)									
GPT-4V [2]	Base	3.49	0.28	13.6	7.3	4.6	67.1	30.7	2.6
LLaVA-v1.5 -7B[25]	HACL [17]	2.13	0.50	-	-	-	-	-	-
	POVID [60]	2.08	0.56	48.1	24.4	-	-	-	-
	EOS [57]	2.03	0.59	40.3	17.8	5.1	49.1	22.7	2.0
	HA-DPO [59]	1.97	0.60	39.9	19.9	6.7	49.8	30.9	3.3
	HALVA [36]	2.25	0.54	-	-	6.6	53.0	32.2	3.4
Comparable Results									
LLaVA-v1.5 -7B [25]	Base	1.59	0.72	25.0	9.2	8.0	44.5	31.0	2.2
	+ LCD [29]	-	-	61.0	16.1	-	-	-	-
	+ ICD [48]	-	-	47.4	13.9	-	-	-	-
	+ OPERA [15]	2.15	0.54	45.1	22.3	-	-	-	-
	+ VCD [19]	1.96†	0.64†	23.6†	8.4†	6.7	46.5	27.8	2.0
	+ M3ID [11]	2.14†	0.61†	23.2†	7.3†	6.0	48.9	26.0	1.5
	+ AVISC [51]	2.19†	0.59†	22.1†	7.8†	6.3	46.6	25.6	2.0
	+ Octopus(Ours)	**2.61**	**0.50**	**20.8**	**6.6**	**4.8**	**49.2**	**23.4**	**1.2**
Instruct-BLIP [7]	Base	1.84	0.64	0.7	9.1	8.4	46.4	31.1	2.6
	+ LCD [29]	-	-	17.4	10.7	-	-	-	-
	+ ICD [48]	-	-	15.2	8.0	-	-	-	-
	+ OPERA [15]	-	-	16.6	6.8	-	-	-	-
	+ VCD [19]	1.75†	0.64†	0.8†	8.9†	7.6	47.7	29.9	2.2
	+ M3ID [11]	1.70†	0.65†	0.9†	7.6†	6.9	47.2	27.5	2.2
	+ AVISC [51]	2.03†	0.59†	0.7†	8.3†	6.7	46.7	28.0	2.6
	Octopus (Ours)	**2.31**	**0.49**	**0.5**	**6.8**	**6.1**	**48.5**	**22.2**	**1.3**

Table 2. Comparison with the state-of-the-art methods for the discriminative tasks across two datasets. † signifies results reproduced with the official implementation codes.

	AMBER		POPE_MSCOCO							
	Discrimination		Random		Popular		Adversarial		ALL	
	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1
LLaVA-1.5-7B[25]	67.00	71.10	83.77	81.94	82.57	80.86	79.77	78.47	82.04	80.42
+ICD [48]	-	-	**87.51**	83.28	83.15	83.91	79.13	80.41	83.26	82.53
+ConVis [32]	-	-	84.70	-	83.20	-	81.10	-	83.00	-
+OPERA [15]	-	-	84.40	-	83.40	-	81.20	-	83.00	-
+VCD [19]	67.30	71.10	85.43	83.99	83.17	81.94	80.27	79.49	82.96	81.81
+M3ID [11] †	67.25	70.90	86.13	81.85	82.07	80.77	79.50	78.15	82.57	80.26
+AVISC [51]	70.70	75.45	84.67	82.21	83.67	81.27	81.83	79.55	83.39	81.01
+Octopus (Ours)	**76.70**	**82.70**	**87.51**	**85.40**	**85.20**	**84.19**	**82.22**	**81.44**	**85.79**	**83.44**
InstructBLIP [7]	68.20	74.60	81.53	81.19	78.47	78.75	77.43	78.00	79.14	79.31
+ICD [48]	-	-	84.36	83.82	77.88	78.70	75.17	77.23	79.14	79.92
+OPERA [15]	-	-	84.57	83.74	78.24	79.15	74.59	76.33	79.13	79.74
+VCD [19]	69.65	75.90	82.03	81.56	79.13	79.20	77.23	77.72	79.46	79.49
+M3ID [11] †	69.05	75.25	82.33	81.53	80.90	80.42	78.53	78.49	80.59	80.15
+AVISC [51]	72.60	78.60	86.03	84.41	84.27	82.77	81.83	80.67	84.04	82.62
+Octopus (Ours)	**74.00**	**79.70**	**86.63**	**85.30**	**84.90**	**83.55**	**82.83**	**81.43**	**84.79**	**83.43**

Table 3. **Ablation study.** We select different numbers of contrastive decoding methods as candidates to demonstrate the effectiveness of our approach. "Str1, Str2, Str3" indicate CD strategy VCD [19], M3ID [11] and AVISC [51] in Sec. 3.2, respectively.

	Str1	Str2	Str3	Octopus	CHAIR ↓	Cover. ↑	Hal. ↓	Cog. ↓
1					8.0	44.5	31.0	2.2
2	✓	✓	✓		6.9	46.2	26.1	2.2
3	✓	✓		✓	5.5	48.7	25.8	1.5
4	✓		✓	✓	5.7	48.2	25.3	1.5
5		✓	✓	✓	5.5	48.4	26.2	1.6
6	✓	✓	✓	✓	**4.8**	**49.2**	**23.4**	**1.2**

Table 4. The effects of different settings for our method.

		CHAIR ↓	Cover. ↑	Hal. ↓	Cog. ↓
1	Cover	5.4	**50.1**	26.1	1.2
2	Average	5.0	49.8	23.0	1.4
3	Monte Carlo [49]	5.1	48.1	23.9	1.3
4	PPO [37]	5.8	47.5	**23.0**	1.5
5	Pooling [6]	6.2	44.8	25.0	1.6
6	Cross-atteition [43]	5.1	49.0	24.2	1.3
7	Ours	**4.8**	49.2	23.4	**1.2**

can mitigate hallucinations to some extent. In 3-5 rows, we show the results of using two different contrastive decoding methods with our Octopus. We find that our Octopus can provide more accurate outputs compared to the random selection. Finally, when we apply three CD strategies together, the hallucination contents are eliminated to a great extent. More importantly, the above experiments demonstrate that our framework has high expansibility and the performance can be improved by introducing more "tentacles".

5.4. Different Model Settings

In Table 4, we explore several alternative settings to further discuss the proposed method on the AMBER dataset.

Different Criterion. Considering that different fields prioritize distinct metrics. For instance, the recommendation domain emphasizes recall [12], while factual precision is more critical in the medical field [3]. Therefore, we focus on verifying the flexibility of our method in the first two rows. In particular, we utilize the Cover score [47] and average score (*i.e.,* averaged by Cover and CHAIR) as metrics to divide the workflow $\mathcal{A}$, respectively. The results show that our Octopus can be easily modified to adapt to the specific needs of different domains.

Different Reinforcement Learning Algorithms. In rows 3-4, we apply different RL-based algorithms to optimize our model. Specifically, we employ Monte-Carlo sampling [49] and PPO [37] to train our Octopus, respectively. It can be seen that these alternative reinforcement learning methods can also achieve satisfactory performance.

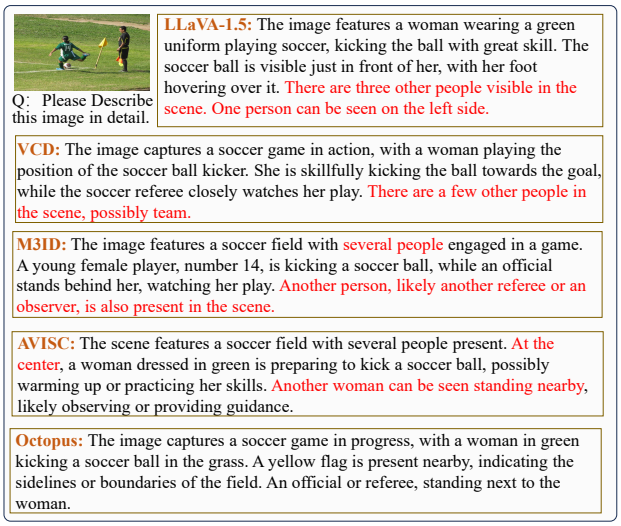

Figure 6. Comparison of generated description with different CD strategies and our method. The hallucination contents are depicted by red text color.

Different Model Architectures. In rows 5-6, we conduct experiments with different model architectures. We replace the vanilla self-attention structure with max pooling [6] and cross-attention [43], respectively. The results show that these transformer-based structures achieve better performance due to their enhanced modeling capabilities.

5.5. Qualitative Evaluation.

To further analyze and verify the proposed method, we visualize the qualitative results on AMBER dataset. As shown in Fig. 6, given an image and a language instruction, we present the responses of our Octopus and other methods including LLaVA-1.5 [25], VCD [19], M3ID [11], AVISC [51]. The red color is used to highlight these hallucinated words. Compared to the original outputs or single CD methods, our Octopus can better eliminate hallucinations and provide a more accurate understanding of the given image.

6. Conclusion

In this paper, we first explore the mechanism behind hallucination occurrences. Extensive experiments show that hallucination causes are hybrid and each generative step suffers from different challenges. Based on the above insights, we propose a new framework Octopus that can adaptively classify hallucination types and build dynamic workflows for different inputs. More importantly, the Octopus has excellent deployability and expansibility, making it a versatile tool for various fields. We expect that this work can provide a general framework to alleviate hallucination challenges across different scenarios.

References

[1] David Abel, David Hershkowitz, and Michael Littman. Near optimal behavior via approximate state abstraction. In *International Conference on Machine Learning*, pages 2915–2923. PMLR, 2016. 5

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 6, 7

[3] Suriya Ganesh Ayyamperumal and Limin Ge. Current state of llm risks and ai guardrails. *arXiv preprint arXiv:2406.12934*, 2024. 8

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. 2023. 1, 2

[5] Mary Beckwith and Frank Restle. Process of enumeration. *Psychological Review*, 73(5):437, 1966. 2, 4

[6] Y-Lan Boureau, Jean Ponce, and Yann LeCun. A theoretical analysis of feature pooling in visual recognition. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 111–118, 2010. 8

[7] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: towards general-purpose vision-language models with instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2024. Curran Associates Inc. 1, 2, 6, 7

[8] Xuerui Dai. Hybridnet: A fast vehicle detection system for autonomous driving. *Signal Processing: Image Communication*, 70:79–88, 2019. 1

[9] Long Hoang Dang, Thao Minh Le, Vuong Le, Tu Minh Phuong, and Truyen Tran. Sadl: An effective in-context learning method for compositional visual qa. *arXiv preprint arXiv:2407.01983*, 2024. 2

[10] Deng-Ping Fan, Tao Zhou, Ge-Peng Ji, Yi Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Inf-net: Automatic covid-19 lung infection segmentation from ct images. *IEEE transactions on medical imaging*, 39(8):2626–2637, 2020. 1

[11] Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14303–14312, 2024. 2, 3, 4, 5, 6, 7, 8

[12] Asela Gunawardana and Guy Shani. A survey of accuracy evaluation metrics of recommendation tasks. *Journal of Machine Learning Research*, 10(12), 2009. 8

[13] Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18135–18143, 2024. 1, 2, 3

[14] MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, 51(6):1–36, 2019. 1, 2

[15] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multimodal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427, 2024. 7

[16] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 3

[17] Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27036–27046, 2024. 7

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[19] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882, 2024. 2, 3, 4, 5, 6, 7, 8

[20] Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. Silkie: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665*, 2023. 6

[21] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 3, 6

[22] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yanfeng Wang, and Cewu Lu. Transferable interactiveness knowledge for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3585–3594, 2019. 1

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 3, 6

[24] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2023. 1, 2, 3

[25] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Pro-*

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 26296–26306, 2024. 1, 2, 3, 6, 7, 8

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

[27] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2

[28] Jiaying Lu, Jinmeng Rao, Kezhen Chen, Xiaoyuan Guo, Yawen Zhang, Baochen Sun, Carl Yang, and Jie Yang. Evaluation and enhancement of semantic grounding in large vision-language models. In *AAAI-ReLM Workshop*, 2024. 1, 2, 3

[29] Avshalom Manevich and Reut Tsarfaty. Mitigating hallucinations in large vision-language models (lvlms) via language-contrastive decoding (lcd). *arXiv preprint arXiv:2408.04664*, 2024. 2, 7

[30] Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734*, 2024. 6

[31] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. 6

[32] Yeji Park, Deokyeong Lee, Junsuk Choe, and Buru Chang. Convis: Contrastive decoding with hallucination visualization for mitigating hallucinations in multimodal large language models. *arXiv preprint arXiv:2408.13906*, 2024. 7

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[34] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 5

[35] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018. 3, 4, 6

[36] Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Ö Arık, and Tomas Pfister. Mitigating object hallucination via data augmented contrastive tuning. *arXiv preprint arXiv:2405.18654*, 2024. 6, 7

[37] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 2, 5, 6, 8

[38] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer, 2022. 3

[39] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv e-prints*, pages arXiv–2309, 2023. 3, 6

[40] Wei Suo, Mengyang Sun, Peng Wang, Yanning Zhang, and Qi Wu. Rethinking and improving feature pyramids for one-stage referring expression comprehension. *IEEE Transactions on Image Processing*, 32:854–864, 2022. 1

[41] Wei Suo, Mengyang Sun, Weisong Liu, Yiqi Gao, Peng Wang, Yanning Zhang, and Qi Wu. S3c: Semi-supervised vqa natural language explanation via self-critical learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2646–2656, 2023. 1

[42] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5, 8

[44] Bin Wang, Fan Wu, Xiao Han, Jiahui Peng, Huaping Zhong, Pan Zhang, Xiaoyi Dong, Weijia Li, Wei Li, Jiaqi Wang, et al. Vigc: Visual instruction generation and correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5309–5317, 2024. 1, 2, 3

[45] Fei Wang, Wenxuan Zhou, James Y Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. mdpo: Conditional preference optimization for multimodal large language models. *arXiv preprint arXiv:2406.11839*, 2024. 3, 6

[46] Guankun Wang, Long Bai, Wan Jun Nah, Jie Wang, Zhaoxi Zhang, Zhen Chen, Jinlin Wu, Mobarakol Islam, Hongbin Liu, and Hongliang Ren. Surgical-lvlm: Learning to adapt large vision-language model for grounded visual question answering in robotic surgery. *arXiv preprint arXiv:2405.10948*, 2024. 2

[47] Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023. 3, 4, 6, 8

[48] Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. *arXiv preprint arXiv:2403.18715*, 2024. 2, 7

[49] Yi Wang, Kok Sung Won, David Hsu, and Wee Sun Lee. Monte carlo bayesian reinforcement learning. *arXiv preprint arXiv:1206.6449*, 2012. 2, 5, 6, 8

[50] Xueru Wen, Xinyu Lu, Xinyan Guan, Yaojie Lu, Hongyu Lin, Ben He, Xianpei Han, and Le Sun. On-policy fine-grained knowledge feedback for hallucination mitigation. *arXiv preprint arXiv:2406.12221*, 2024. 2

[51] Sangmin Woo, Donguk Kim, Jaehyuk Jang, Yubin Choi, and Changick Kim. Don't miss the forest for the trees: Attentional vision calibration for large vision language models. *arXiv preprint arXiv:2405.17820*, 2024. 2, 3, 4, 5, 6, 7, 8

[52] Yutong Xie, Jianpeng Zhang, Yong Xia, Michael Fulham, and Yanning Zhang. Fusing texture, shape and deep model-learned information at decision level for automated classification of lung nodules on chest ct. *Information Fusion*, 42: 102–110, 2018. 1

[53] Jing Xu, Andrew Lee, Sainbayar Sukhbaatar, and Jason Weston. Some things are more cringe than others: Preference optimization with the pairwise cringe loss. *arXiv preprint arXiv:2312.16682*, 2023. 6

[54] Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12944–12953, 2024. 2

[55] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816, 2024. 2

[56] Tongtian Yue, Jie Cheng, Longteng Guo, Xingyuan Dai, Zijia Zhao, Xingjian He, Gang Xiong, Yisheng Lv, and Jing Liu. Sc-tune: Unleashing self-consistent referential comprehension in large vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13073–13083, 2024. 2

[57] Zihao Yue, Liang Zhang, and Qin Jin. Less is more: Mitigating multimodal hallucination from an eos decision perspective. *arXiv preprint arXiv:2402.14545*, 2024. 6, 7

[58] Lijun Zhang, Wei Suo, Peng Wang, and Yanning Zhang. A plug-and-play method for rare human-object interactions detection by bridging domain gap. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8613–8622, 2024. 2

[59] Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*, 2023. 6, 7

[60] Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*, 2024. 7

[61] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1, 2