HalluciDoctor: Mitigating Hallucinatory Toxicity in Visual Instruction Data

Qifan Yu[1] Juncheng Li[1†] Longhui Wei[2] Liang Pang[3] Wentao Ye[1] Bosheng Qin[1]
Siliang Tang[1] Qi Tian[2] Yueting Zhuang[1]
[1]Zhejiang University, [2]Huawei Cloud, [3] Institute of Computing Technology, Chinese Academy of Sciences
{yuqifan, junchengli, bsqin, siliang, yzhuang}@zju.edu.cn
{weilonghui1, tian.qi1}@huawei.com, pangliang@ict.ac.cn

Abstract

Multi-modal Large Language Models (MLLMs) tuned on machine-generated instruction-following data have demonstrated remarkable performance in various multi-modal understanding and generation tasks. However, the hallucinations inherent in machine-generated data, which could lead to hallucinatory outputs in MLLMs, remain under-explored. This work aims to investigate various hallucinations (i.e., object, relation, attribute hallucinations) and mitigate those hallucinatory toxicities in large-scale machine-generated visual instruction datasets. Drawing on the human ability to identify factual errors, we present a novel hallucination detection and elimination framework, ***HalluciDoctor****, based on the cross-checking paradigm. We use our framework to identify and eliminate hallucinations in the training data automatically. Interestingly,* ***HalluciDoctor*** *also indicates that spurious correlations arising from long-tail object co-occurrences contribute to hallucinations. Based on that, we execute counterfactual visual instruction expansion to balance data distribution, thereby enhancing MLLMs' resistance to hallucinations. Comprehensive experiments on hallucination evaluation benchmarks show that our method successfully mitigates 44.6% hallucinations relatively and maintains competitive performance compared to LLaVA. The data and code for this paper are publicly available.*[1]

1. Introduction

Recently, Multi-modal Large Language Models (MLLMs) have made promising progress on multi-modal tasks, such as image captioning, visual question-answering, and visual conversations [1, 12, 13, 31]. Additionally, inspired by the impressive instruction-following capability of LLMs [4, 28, 30], several more powerful MLLMs [5, 14, 21, 36, 44] have recently emerged, extending instruction-tuning to the multi-modal space. Due to the scarcity of visual-language instruction-following data, recent research[21, 44] presents a data reformation approach, which leverages text-only LLMs conditioned on image captions and bounding boxes to create instruction-following data involving visual content. However, these visual instructions might include hallucinatory responses incongruent with the image content, as they are produced by text-only large language models. For instance, as shown in Figure 1 (a), the response in the visual instruction data includes object hallucinations such as "a few traffic lights" and relation hallucinations like "people gathered for flying kites" instead of "sitting and watching". These hallucinatory responses may compromise the MLLM's ability to perceive the real world accurately.

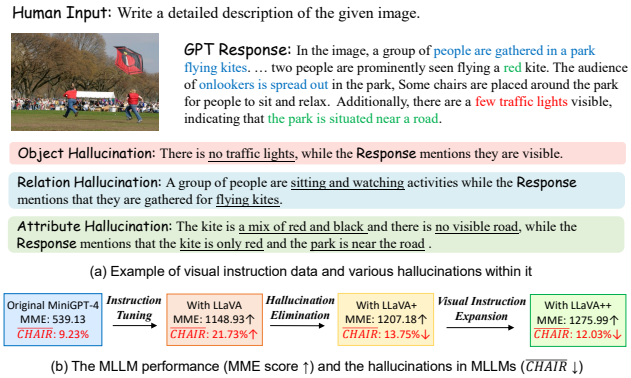


Figure 1. (a) On the top, we show an example of visual instruction and various hallucinatory toxicities within it. (b) At the bottom, we show that refined LLaVA++ from HalluciDoctor can alleviate hallucinatory toxicity to MLLM and improve its performance.

Motivated by this insight, we systematically definite various kinds of hallucinations (*i.e.*, object, relation, attribute hallucinations) and investigate them in visual instruction datasets. Using the widely used dataset LLaVA-Instruction-158K [21], we construct extended CHAIR metrics to comprehensively evaluate the impact of these visual instructions on modern MLLMs, considering both performance and hallucination issues. While instruction-tuning on LLaVA data

†Juncheng Li is the corresponding author.
[1]https://github.com/Yuqifan1117/HalluciDoctor

improves MLLM performance, it significantly increases the probability of producing hallucinations (Figure 1 (b), original MiniGPT-4 v.s. with LLaVA finetuning). These observations confirm with recent studies [7, 15, 20] that current machine-generated instruction-following data contains massive pernicious hallucinations (32.6% in LLaVA) that cause MLLM to produce inaccurate outputs. This pushes us to focus on mitigating those hallucinatory toxicities.

Previous works mainly focus on collecting extra remedial training data [20, 23, 32] or utilize additional plug-in models [7, 18, 38] to mitigate the hallucinations during inference. LRV-Instruction [20] proposes to add extra negative instructions to increase the robustness of the MLLM against hallucinations. M-HalDetect [7] mitigates hallucinations by incorporating an additional trained reward model during the inference phase. However, these methods either raise training labor costs or prolong inference time. Moreover, they merely superficially suppress the hallucinatory output of MLLMs, largely neglecting the inherent hallucinatory toxicity in the visual instruction dataset, which causes the hallucinatory errors in existing MLLMs. This leads to sub-optimal hallucination elimination for MLLMs.

In contrast to the methods above, we aim to eradicate hallucinations in machine-generated visual instruction data. The primary challenge is how to accurately detect and remove various hallucinations from massive such data without manual annotations. For this, we propose a flexible Hallucination Detection and Elimination Framework, namely **HalluciDoctor**, which automatically detects various hallucinations in arbitrary positions and dispels them based on a cross-checking paradigm. The key insight is that when asked about the hallucinatory content of a given image, the responses from different MLLM experts typically tend to vary and can even contradict each other. Specifically, as shown in Figure 2, HalluciDoctor breaks the hallucination detection procedure into three sub-processes: 1) *Answer Chunks Extraction*: extract all answer chunks including objects, relations, and attributes by textual scene graph parsing as description-oriented answers; 2) *Answer-based Question Generation*: generate corresponding fine-grained questions with diverse types for each answer; 3) *Consistency Cross-Checking*: obtain image-oriented candidate answers from multiple MLLMs and cross-check the consistency between description-oriented answer chunks and their corresponding image-oriented answers. Subsequently, HalluciDoctor identifies those semantic chunks with consistency scores below a threshold as hallucinatory chunks. It eliminates these hallucination errors without disrupting the contextual semantics, resulting in the rectified dataset, LLaVA+. This significantly alleviates hallucinations in MLLMs (Figure 1 (b), with LLaVA v.s. with LLaVA+).

In our exploration of eliminating hallucinations in visual instruction data, we find that HalluciDoctor not only assists in locating hallucinations but also indicates the spurious correlations causing them, which stem from the long-tail distribution of object co-occurrences. These spurious correlations can mislead MLLMs into erroneously inferring the presence of objects that do not exist in the images. Inspired by concepts of counterfactual generation [10, 16, 35, 41], we propose a seesaw-based strategy for counterfactual visual instruction expansion. It resolves this issue by balancing the long-tail object co-occurrence distribution through two collaborative factors, ultimately creating a more robust visual instruction dataset, LLaVA++. This enables MLLMs to concentrate on accurately perceiving the content of images instead of spurious associations, thereby strengthening their resistance to hallucinations and overall performance (Figure 1 (b), with LLaVA+ v.s. with LLaVA++).

Our main contributions are summarized as follows:

- To the best of our knowledge, we are the first to comprehensively investigate the severe hallucination toxicity in existing machine-generated visual instruction datasets.
- We propose a novel **HalluciDoctor** method to detect various hallucinations by a consistency cross-checking paradigm and dispel them in a low-resource way.
- Based on HalluciDoctor, we further automatically generate more counterfactual instruction data to improve the MLLMs' resistance to hallucinations.
- Our empirical study confirms our method's effectiveness in eliminating hallucinations in visual instruction data and improving MLLMs' robustness.

2. Related works

2.1. Multi-modal Large Language Model

With the remarkable generalizability of LLMs in a zero-shot setting [26, 33, 34, 42], early works integrating LLMs with visual modality have demonstrated impressive visual-language understanding ability [8, 11, 13, 31, 39]. Recently, more powerful MLLMs [2, 5, 14, 25, 36, 37, 44] have emerged to mimic human perceptual capabilities for unseen vision-language tasks. Generally, MLLMs align the vision encoder into the LLM by a cross-modal alignment network (*e.g.*, a linear projection layer in MiniGPT-4 [44], a visual abstractor in mPLUG-Owl [36], and Q-former in InstructBLIP [5]). The training process of MLLMs mainly contains two stages: the first pre-training and the second multi-modal instruction tuning. Moreover, LLaVA [21] leverages powerful LLMs to obtain extensive visual instruction data, paving the way for acquiring the instruction-following ability of MLLMs. This is essential for constructing more powerful MLLMs in a low-resource way.

2.2. MLLM Hallucination

Although MLLMs have demonstrated remarkable performances in various VL tasks, they still suffer from the hallu-

cination phenomenon that textual outputs conflict with the visual content. Current research on MLLM hallucinations mainly focuses on the detection and elimination of hallucinations [7, 15, 18, 20, 32, 43]. [15] solely concentrates on object hallucinations and treats hallucination detection as a binary classification issue, limiting its evaluation for open-ended responses. HalDetect [7] identifies hallucinations by training a specialized classifier. But both methods need manual ground-truth answer collection and only consider the simplest object hallucinations. For hallucination mitigation within MLLMs, previous works generally either collect more high-quality data manually [20] or attach an extra correction model [7, 18, 32]. Furthermore, existing works solely concentrate on direct hallucinations in MLLM reasoning, ignoring the essential hallucinatory toxicity in the visual instruction data itself. Contrary to the methods above, we shift our attention to diverse hallucinations in the visual instruction data and devise an automated framework to detect and eliminate potential hallucinatory toxicity.

3. The Toxicity of Visual Instruction Data

Due to the scale limitation of multi-modal instruction-following data, there is a growing interest in self-generated instructions for MLLMs [21, 44]. However, since these visual instructions are generated by text-only GPT-4, they may contain numerous hallucinations, leading MLLMs to produce responses that inaccurately represent the images. To our knowledge, it is the first work to systematically analyze the hallucinatory toxicity of visual instruction datasets.

3.1. Hallucination Metrics

To better analyze hallucinatory toxicity in the dataset, we first categorize three types of hallucinations frequently appearing in it: 1) *object hallucination* is the object that appears in the description but not in the image. 2) *relation hallucination* involves the relation between corresponding objects that exhibits inconsistency between descriptions and images. 3) *attribute hallucination* refers to inaccurate object properties in the description, such as size, color, and states.

The current popular metric CHAIR [29] only calculates the proportion of nonexistent objects in the description. Thus, we extend the naive CHAIR metric into more complex scenarios to evaluate various hallucinations. Initially, we incorporate synonym lists into annotated objects and phrases, forming an enhanced ground truth set. Subsequently, we split the description into sentences and extracted all objects, relations, and attributes for a comprehensive assessment. The extended CHAIR metric then calculates the ratio of sentences containing hallucinatory elements not present in the image. Accordingly, the definition of extended CHAIR, including CHAIR_{obj}, CHAIR_{rel}, and CHAIR_{attri}, is delineated as follows:

Dataset	#Samples	$\text{CHAIR}_{obj}\downarrow$	$\text{CHAIR}_{rel}\downarrow$	$\text{CHAIR}_{attri}\downarrow$	Length
LLaVA [21]	158K	28.1	36.0	33.7	96.1
LLaVA+	158K	**8.3**	**20.7**	**17.1**	87.8
MiniGPT4-Instruction [44]	3.5K	22.6	35.6	31.6	70.8
MiniGPT4-Instruction+	3.5K	**13.3**	**21.7**	**23.8**	61.8

Table 1. The statistics of three types of hallucinations in visual instruction datasets and comparison with their corresponding rectified version by HalluciDoctor (**bolded rows**).

$$\text{CHAIR}_{obj} = \frac{|\{\text{sentences with nonexistent object}\}|}{|\{\text{all sentences}\}|} \quad (1)$$

$$\text{CHAIR}_{rel} = \frac{|\{\text{sentences with nonexistent relation}\}|}{|\{\text{all sentences}\}|} \quad (2)$$

$$\text{CHAIR}_{attri} = \frac{|\{\text{sentences with nonexistent attribute}\}|}{|\{\text{all sentences}\}|} \quad (3)$$

The higher CHAIR score indicates there exist more hallucinations in the description. Notably, we compute the CHAIR_{rel} and the CHAIR_{attri} only among existent objects to avoid misjudging compositional errors influenced by object hallucinations as additional hallucination errors.

3.2. Hallucinatory Toxicity Statistics

Utilizing our extended CHAIR metric designed for in-depth hallucination analysis, we meticulously examine the hallucination frequency within machine-generated visual instruction data. We concentrate on prevalent, machine-generated visual instruction datasets, namely LLaVA [21] and MiniGPT4-Instruction [44]. LLaVA consists of 158K distinct instruction-following samples generated by GPT-4 [25], while MiniGPT4-Instruction includes about 3.5K instances refined by ChatGPT [24] from detailed descriptions. To tackle the challenge of incomplete annotations in the above datasets, particularly regarding relations and attributes, we adopt GroundingDINO [22] to annotate objects and use the image-text similarity of BLIP [12] to judge the existence of relations and attributes. By incorporating these pseudo-labels, we thoroughly assess the hallucinatory toxicity in those datasets, as depicted in Table 1. It can be seen that machine-generated visual instruction data uniformly exhibit various distinct types of hallucinations. However, the proxy detection approach lacks flexibility in addressing a variety of unidentified hallucinations due to its reliance on accurate annotation. Therefore, we propose a general hallucination detection and elimination framework **HalluciDoctor** based on the consistency cross-checking paradigm to handle the potential hallucinations in the training data.

4. HalluciDoctor Framework

As illustrated in Figure 2, our HalluciDoctor framework consists of two primary modules. **Hallucination Cross-**

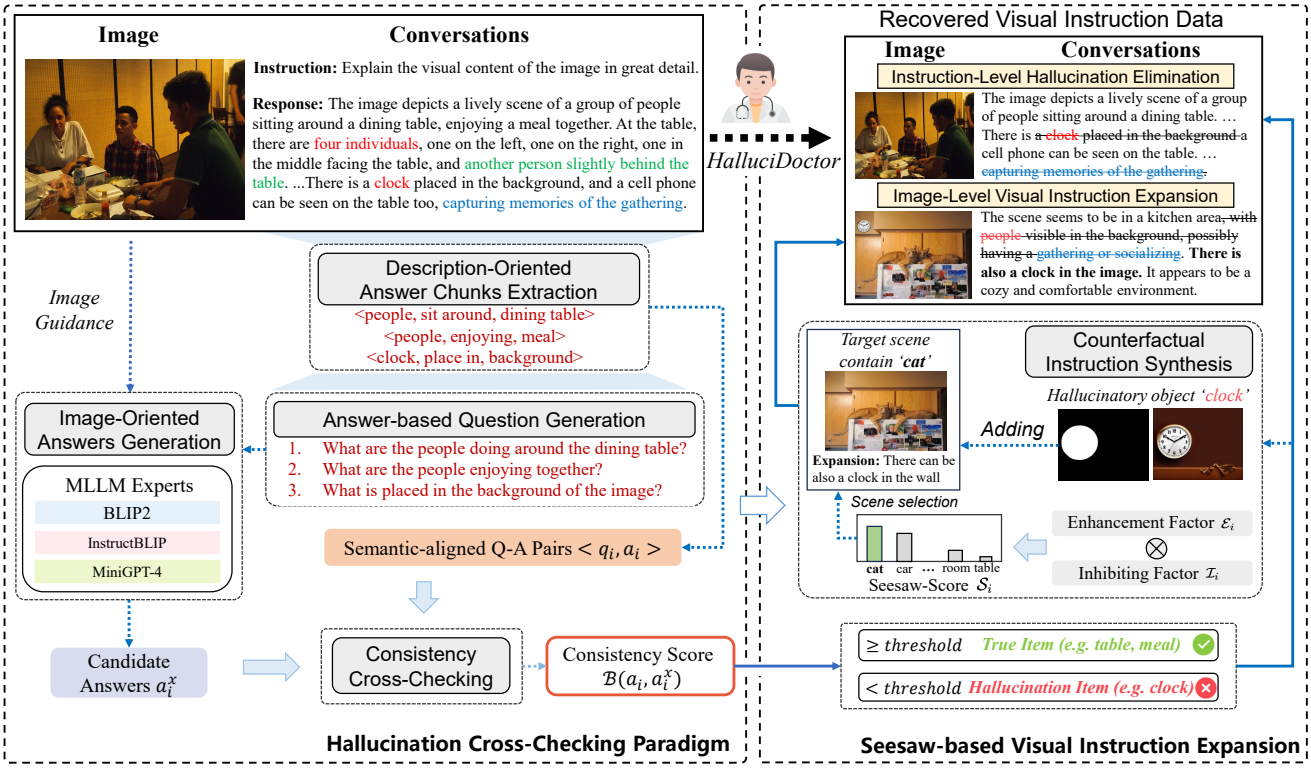


Figure 2. Overview of our proposed HalluciDoctor for automatically eliminating hallucinatory toxicity in visual instruction data and enhancing MLLM's resistance to hallucinations. We summarize the process into four steps: (1) HalluciDoctor first extracts description-oriented answers for semantic analysis and formulates corresponding questions. (2) Image-oriented candidate answers for these questions are then gathered from various MLLMs. (3) HalluciDoctor will identify and remove hallucinatory chunks via consistency cross-checking. (4) Lastly, It creates counterfactual instructions guided by preceding steps to expand the dataset and mitigate hallucinations radically.

Check Paradigm is designed to probe and eliminate hallucinatory errors in the original visual instruction data. **Seesaw-based Visual Instruction Expansion** produces additional counterfactual visual instructions to reduce the hallucinatory effects caused by spurious correlations and strengthen the MLLMs' resistance to hallucinations.

4.1. Hallucination Cross-Checking Paradigm

Overview. To exhaustively identify various types of hallucinations (*i.e.*, object hallucinations, relation hallucinations, attribute hallucinations) in the corresponding description of each image, we introduce a hallucination cross-checking paradigm. Our insight is to decompose this demanding detection task into several simpler answer consistency-checking tasks. This paradigm comprises three subtasks: answer chunks generation, question generation, and consistency cross-checking. Then, HalluciDoctor eliminates any detected hallucinations. More details are in Appendix B.

Answer Chunks Generation. Since the generated description contains massive concepts, we employ a textual scene graph parser [17] to extract description-oriented answer chunks to represent the specific semantics, including objects, attributes, and relations. Given an image I with its instruction-following data $(\mathrm{X_q}, \mathrm{X_a})$, where $\mathrm{X_q}$ is the instruction from human and $\mathrm{X_a}$ is the generated description from GPT-4. We extra all answer chunks from instruction-following data $(\mathrm{X_q}, \mathrm{X_a})$ together as follows,

$$A = \{a_1, ..., a_n\} \tag{4}$$

where a_i represents the i^{th} answer chunk in the $(\mathrm{X_q}, \mathrm{X_a})$.

Question Generation. Upon generating answer chunks, we construct corresponding questions, which are then used to derive image-oriented candidate answers. We employ LLM, like ChatGPT as potent question generators to encompass the wide diversity of answer chunks and question types. Thus, we can generate questions tailored to different answers, including concrete objects, abstract relations, and attribute descriptions as follows,

$$Q = \{q_1, ..., q_n\} \tag{5}$$

where q_i represents the i^{th} question corresponding to a_i.

Consistency Cross-Checking. The consistency cross-checking step aims to verify the consistency between each answer chunk a_i and its corresponding content in the image. To obtain the specific image content, we use the generated

questions to derive image-oriented candidate answers based on MLLM experts (*e.g.*, BLIP2 [13], InstructBLIP [5], and MiniGPT-4 [44]). Formally, let $\{\mathcal{F}_x\}$ denote an MLLM expert, the prompt template slotted with the reference image I and one question q_i is fed to the MLLM $\mathcal{F}_x$ to produce the image-oriented candidate answer $a_i^x = \mathcal{F}_x(I, q_i)$. Once we collect image-oriented candidate answers $\{a_i^x\}$ for each answer chunk a_i, we will compare their consistency. We employ a Bert-based metric (*i.e.*, BEM [3]) to evaluate their consistency since this metric provides more flexibility in the answer formulation than strictly hard matching. Let $\mathcal{B}(\cdot, \cdot|q)$ denote the BEM score between two answers according to the question, the final *ConScore* of the answer chunk a_i is calculated by voting as follows,

$$ConScore_i = \frac{1}{m}\sum_{x=1}^{m}\mathcal{B}(a_i, a_i^x) \tag{6}$$

where m denotes the number of MLLM experts. We will consider the answer chunk a_i as a hallucination when its ConScore < 0.5. More discussion of BEM evaluation and threshold determination is presented in Appendix B.2.

Hallucination Elimination. Benefiting from the above subtasks, we can accurately locate hallucinatory chunks with their contexts. We employ ChatGPT to automatically remove the hallucinatory chunks based on the context and guarantee the coherence and harmony of the corresponding sentences. After this rectification process, we obtain more accurate visual instruction data by diminishing erroneous hallucinatory descriptions, denoted as LLaVA+. This step is designed to alleviate the hallucinatory toxicity in machine-generated visual instruction data for MLLM training.

4.2. Seesaw-based Visual Instruction Expansion

In our efforts to locate hallucinations within visual instruction data, we have observed a notable pattern: these hallucinations frequently occur alongside objects that often appear together due to the long-tail distribution of object co-occurrences. For instance, in visual datasets, commonly co-occurring objects like 'cars' and 'roads' may inadvertently lead to spurious correlations, resulting in hallucinations. Consequently, a model trained on such data might incorrectly infer the presence of a 'car' when encountering a 'road' even in images where such a pairing is absent. This misunderstanding is the root cause of many hallucinations in visual data. By integrating hallucinatory objects into tail scenes where they rarely appear, we introduce counterfactual interventions [27] to mitigate the spurious correlations among strongly associated objects.

However, selecting such scenes is challenging, as it requires balancing the rarity of co-occurrence with the hallucinatory objects and their contextual plausibility. To this end, we propose a seesaw-based strategy with an enhancement factor and an inhibiting factor to adaptively select target scenes for counterfactual instruction expansion.

Enhancement Factor. Given an illusory object o in the response, we denote the object in the response that appears most frequently with o across all annotations as o^* and their co-occurrence frequency as n^*. The enhancement factor $\mathcal{E}_i$ is designed to increase the weight of other objects o_i that rarely co-occur with o and is computed as follows,

$$\mathcal{E}_i = \begin{cases} \frac{n^*}{\max(n_i,1)} & , \text{ if } n_i \leq n^* \\ 1 & , \text{ if } n_i > n^* \end{cases} . \tag{7}$$

where n_i represents the co-occurrence frequency between the hallucinatory object o and other object o_i, and $\mathcal{E}_i$ is inversely correlated with their co-occurrence frequency.

Inhibiting Factor. Although the enhancement factor effectively promotes the co-occurrence of the hallucinatory object with its infrequently co-occurring objects, it overlooks the contextual plausibility of these co-occurring pairs. Therefore, we introduce the inhibiting factor $\mathcal{I}_i$ to suppress the weight of objects in rare combinations with the contextually relevant object o^* as follows,

$$\mathcal{I}_i = \begin{cases} \frac{m_i}{n^*} & , \text{ if } m_i \leq n^* \\ 1 & , \text{ if } m_i > n^* \end{cases} . \tag{8}$$

where m_i represents the co-occurrence frequency between the o's most contextually relevant object o^* and other objects o_i. Once we obtain $\mathcal{E}_i$ and $\mathcal{I}_i$, we calculate the Seesaw-Score $\mathcal{S}_i$ as demonstrated below.

$$\mathcal{S}_i = \mathcal{E}_i * \mathcal{I}_i . \tag{9}$$

In eq. 7-9, rare objects exhibit a higher $\mathcal{E}$, while reasonable combinations show a higher $\mathcal{I}$. In this way, we identify scenes containing objects with the highest seesaw scores as target scenes and integrate o into these scenes. We phrase its description as `There is also a/an {object} in the image` and create counterfactual instructions. Subsequently, we incorporate hallucinatory objects into suitable locations of target scenes guided by the bounding box, to facilitate the corresponding counterfactual image synthesis. The generated counterfactual instructions are then amalgamated with the rectified dataset, LLaVA+, to form a more robust dataset for MLLM instruction tuning, LLaVA++, thereby reducing the impact of spurious correlations on hallucinations. More details and analyses are in Appendix B.4.

5. Experiments

In this section, we present both qualitative and quantitative experimental results and corresponding analyses to assess HalluciDoctor's superiority. Our focus primarily lies on detailed experiments regarding MLLM hallucinations (§ 5.2) and MLLM performance (§ 5.3), and based

Model Type		Methods	Instance-level			Sentence-level		
			$CHAIR_{obj} \downarrow$	$CHAIR_{rel} \downarrow$	$CHAIR_{attri} \downarrow$	$CHAIR_{obj} \downarrow$	$CHAIR_{rel} \downarrow$	$CHAIR_{attri} \downarrow$
Specific		Faithful Prompt	9.3	11.1	14.1	23.2	24.8	25.5
		LURC[43]	5.7	7.6	13.3	16.0	22.8	28.5
		VIGC[32]	6.1	7.5	11.5	15.2	17.7	22.3
Model-Agnostic	MiniGPT4 (7B)	w/ LLaVA [21]	12.0	12.2	10.1	35.0	34.8	26.3
		w/ LRV [20]	10.0	10.8	13.6	24.9	21.0	24.8
		w/ LLaVA+	5.9	6.1	8.5	19.6	20.5	21.9
		w/ LLaVA++	**5.7**	**5.4**	**7.1**	**16.6**	**17.1**	**20.3**
	mPLUG-Owl (7B)	w/ LLaVA [21]	10.6	10.0	10.3	32.6	32.0	29.1
		w/ LRV [20]	10.3	9.5	13.0	30.8	29.6	32.1
		w/ LLaVA+	7.6	7.1	8.0	22.2	21.1	24.2
		w/ LLaVA++	**6.4**	**5.5**	**6.7**	**19.3**	**17.6**	**16.5**

Table 2. Comprehensive CHAIR evaluation results to show the recovery effect of hallucination elimination by HalluciDoctor for MLLMs.

on these aspects, we organize more comprehensive GPT-4 evaluations (§ 5.4) and human assessments (§ 5.5) to evaluate MLLMs' open-ended capabilities precisely.

5.1. Experimental Setup

Model Setting. In this paper, we utilize the most widely used machine-generated visual instruction data LLaVA-158K [21] to conduct experiments. To provide a comprehensive evaluation, we thoroughly compare our HalluciDoctor (*w/ LLaVA+* and *w/ LLaVA++*, shown in the ■ of Table 2) with various SOTA methods tailored for alleviating hallucinations. We categorize those methods into two categories: (1) Specialized approaches. We incorporate some models requiring additional dedicated modules to mitigate hallucinations, including LURC[43] and VIGC [32], as well as employing explicit faithful prompts to constrain the generation of reliable instruction data (Faithful Prompt). (2) Model-agnostic baselines. They refer to plug-and-play methods for optimization at the dataset level and fine-tune MLLMs on the corresponding instruction data, including *w/ LLaVA* [21] and *w/ LRV* [20]. For model-agnostic baselines, we equip variant datasets with two popular MLLMs: MiniGPT-4 [44] and mPLUG-Owl [36]. We use the official pre-trained MLLM with the image-text alignment stage and only fine-tune it in the second stage for fair comparison.

5.2. Comparion of MLLM Hallucinations

To assess the toxicity of visual instruction data on MLLMs and the efficacy of Hallucidoctor in eliminating hallucinations, we compared MLLMs with Hallucidoctor against baseline models using our extended CHAIR benchmark. In addition to the sentence-level CHAIR evaluation mentioned above (*i.e.*, $CHAIR_S$), we further computed CHAIR at the instance-level (*i.e.*, $CHAIR_I$) by quantifying all nonexistent instances within a sentence to assess the overall distribution of hallucinations. Following previous works [15, 20], we randomly select 500 unique images from the intersection of MSCOCO [19] and Visual Genome [9] for a more detailed evaluation. Notably, these images were different from the ones used in LLaVA-158k and contained various kinds of annotations. Subsequently, we prompt MLLMs with the instruction of `Provide a detailed description of the given image` to generate detailed captions of similar length. We compute $CHAIR_{obj}$, $CHAIR_{rel}$, and $CHAIR_{attri}$ at two different levels and report the results in Table 2. Based on the observation of experimental results, we have summarized the following conclusions:

Visual instruction data has serious hallucinatory toxicity. We compare the MLLM fine-tuned with LLaVA against the original MLLM. The former is more susceptible to generating hallucinations, particularly in object perception. This confirms our previous analysis of the hallucinatory toxicity in visual instruction data, further emphasizing the necessity to eliminate hallucinations therein.

Our HalluciDoctor can be flexibly equipped to different MLLMs for hallucination elimination. We utilize HalluciDoctor to eliminate hallucinations and obtain the LLaVA+. For hallucination evaluation, we integrate this rectified dataset into two backbone models, MiniGPT-4 [44] and mPLUG-Owl [36]. Despite the model diversity, MLLMs with LLaVA+ can consistently reduce the probability of various hallucinations (average reduction of 4.6% / 11.4% in MiniGPT-4 and 2.7% / 8.7% in mPLUG-Owl in two metric levels). The results confirm that our HalluciDoctor, by effectively reducing hallucination errors in visual instruction data (as shown in the bolded rows in Table 1), can alleviate hallucinatory outputs in MLLMs, thereby enhancing their reliability in the real world.

Compared with other model-agnostic methods, our HalluciDoctor outperforms all of them for hallucination elimination. Specifically, MLLMs fine-tuned on LLaVA+ exhibit fewer hallucinations among all three types compared to those trained on the meticulously curated SOTA dataset LRV-Instruction [20], especially for more challenging attribute hallucinations (8.5% v.s. 13.6% in instance-level of MiniGPT-4). Similar results are also observed in the closed-ended POPE evaluation [15], with detailed analysis in Appendix C.2. It indicates that our HalluciDoctor

	Model Type	Methods	Perception	Cognition
MME Benchmark	Specific	Faithful Prompt	696.34	293.57
		LURC[43]	904.61	254.13
		VIGC[32]	879.35	221.79
	MiniGPT4 (7B)	w/ LLaVA [21]	859.64	289.29
		w/ LRV [20]	870.12	291.79
		w/ LLaVA+	889.32	317.86
		w/ LLaVA++	**955.28**	**320.71**
	mPLUG-Owl (7B)	w/ LLaVA	967.34	276.07
		w/ LRV [20]	1008.57	263.93
		w/ LLaVA+	1043.19	282.14
		w/ LLaVA++	**1114.34**	**302.86**

Table 3. Results on MME evaluation [6] of MiniGPT4-7B and mPLUG-Owl-7B. The performance is measured by the sum of the subtasks' scores, where the best score for each partition is bolded.

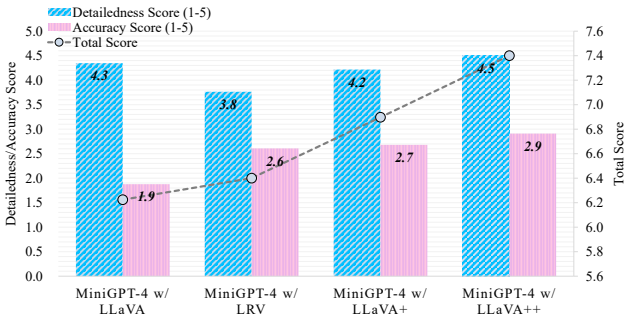


Figure 3. Evaluation scores of detailedness and accuracy for descriptions from MiniGPT-4 with different setups. We visualized the total scores using a gray line, where higher scores indicate more detailed descriptions and fewer hallucinations.

effectively eliminates massive hallucinations in machine-generated data, constructing higher-quality visual instructions to mitigate hallucinatory toxicity in MLLMs.

Visual instruction expansion can effectively reduce hallucinations caused by spurious correlations. With the help of LLaVA++, MLLM obtained fewer object hallucinations than LLaVA+ (17.1% v.s. 20.5% in sentence-level CHAIR_{rel}). It suggests that expanded counterfactual instructions can equalize the long-tail distribution of object co-occurrences, reducing MLLMs' inclination towards incorrect associations. Remarkably, the MLLM fine-tuned on LLaVA++ exhibited the fewest hallucinations across all metrics, highlighting the superiority of HalluciDoctor in enhancing model reliability. To further mitigate the impact of long-tail distributions, it is also promising to incorporate contrastive learning [40] with counterfactual interventions.

5.3. Comparion of MLLM Performance

While the extended CHAIR evaluation affirms HalluciDoctor's efficacy in hallucination elimination, a well-rounded analysis of its impact on MLLM performance remains to be conducted. Therefore, we conduct quantitative analysis on the MME benchmark [6], which evaluates the perception and cognition abilities of MLLMs on 14 subtasks. This setup converts human annotations into a series of "yes or no" questions and measures MLLM performance by calculating the total accuracy score. Table 3 summarizes the cognitive and perceptual performance of MLLMs fine-tuning on different datasets. Compared to LLaVA, LLaVA+ not only mitigates hallucinations but also achieves higher MLLM performance (1207.18 v.s. 1148.93). In comparison to LRV-Instruction [20], LLaVA+ performs better than this SOTA method, even with fewer visual instructions. This indicates that HalluciDoctor still preserves the accurate elements when eliminating hallucinations in visual instructions for better instruction alignment. Additionally, LLaVA++ offers more challenging counterfactual instructions for better generalization. With the aid of LLaVA++, both fine-tuned MLLMs experience further improvement than LLaVA (+127.06 / +173.79 in overall performance).

5.4. GPT-4 Evaluation

Empirically, there is a trade-off between the diversity of descriptions generated by MLLMs and the frequency of hallucinations. To strike a balanced evaluation between hallucinations and performance, we employ GPT-4 for a comprehensive assessment, comparing detailedness and accuracy in descriptions generated by MLLMs. In Figure 3, we visualize the GPT-4 evaluation scores under different MLLM settings. On the one hand, LLaVA significantly enriches descriptive content but at the cost of increased hallucinations. On the other hand, LRV-Instruction [20] reduces such hallucinatory errors at the expense of descriptive diversity, thus constraining the generative potential of the model. In contrast, our HalluciDoctor outperforms other models on a total score, indicating that it can eliminate hallucinations while preserving MLLM's ability to output diverse descriptions.

5.5. Human Evaluation

To more comprehensively assess the open-ended capabilities of MLLMs, we further conduct human evaluations using the OwlEval benchmark. **OwlEval** [36] is another open-ended evaluation set with 82 artificially constructed questions. We quantified responses from all models on the 3-0 scale (aligned with option A-D in the official setting), calculating quality and accuracy scores based on the relevance of the response to the question and the precision of the description, respectively. Additionally, we computed the score variance among each MLLM's responses to evaluate model stability. We show the visualized results for all MLLMs in Figure 5. We observe that MLLMs fine-tuned with LLaVA++ achieve the highest accuracy scores regarding image content while maintaining fidelity to the corresponding questions. This suggests that HalluciDoctor effectively reduces hallucinations without compromising response quality, adeptly addressing a wide range of open-domain questions rather than limiting response length.

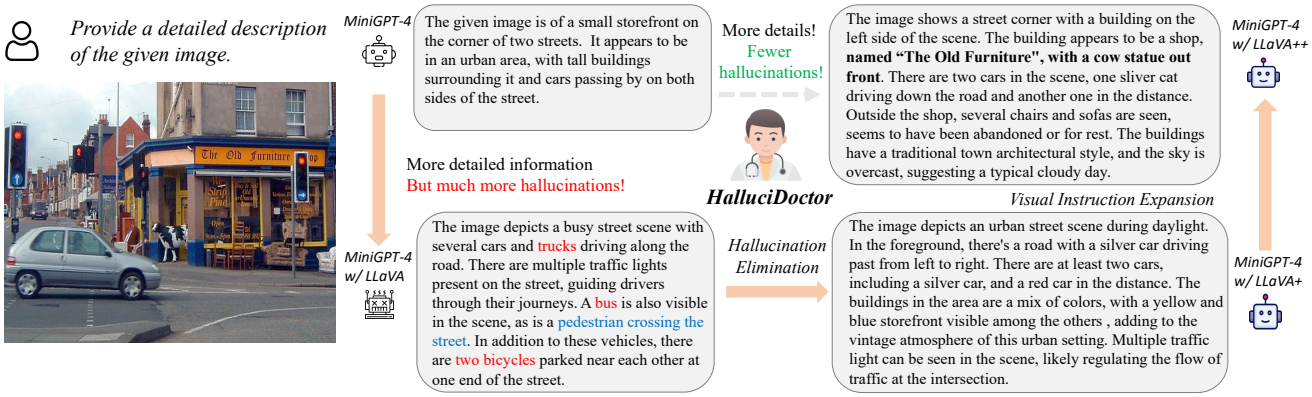
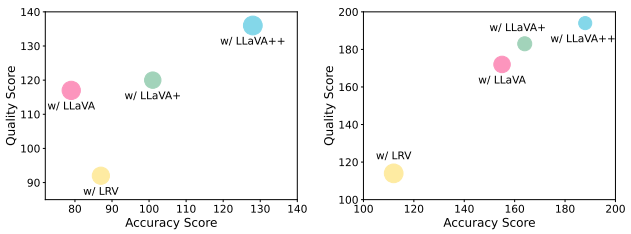


Figure 4. A case study comparing the levels of MLLM hallucination after fine-tuning on various instruction data.

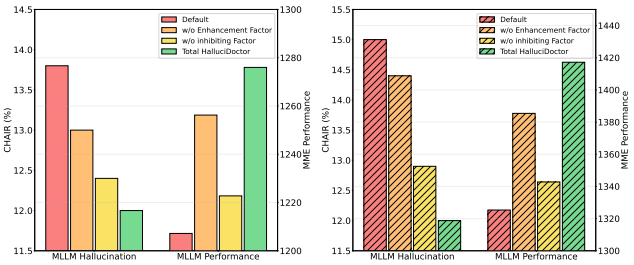


(a) MiniGPT-4 Results (b) mPLUG-Owl Results

Figure 5. Quality score (y-axis, higher is better), accuracy score (x-axis, higher is better), and the stability (circle sizes, smaller is better) of MLLMs' responses on OwlEval benchmark.

Dataset	$\overline{\text{CHAIR}}$ (%)		MME Performance	
	$\text{CHAIR}_I \downarrow$	$\text{CHAIR}_S \downarrow$	Perception ↑	Cognition ↑
MiniGPT4-Instruction [44]	9.2	23.7	616.41	232.71
MiniGPT4-Instruction+	8.4	18.6	659.67	255.03
MiniGPT4-Instruction++	**5.9**	**15.2**	**696.96**	**282.86**

Table 4. CHAIR results and MME performance of applying HalluciDoctor on MiniGPT4-Instruction Dataset [44].



(a) MiniGPT-4 (b) mPLUG-Owl

Figure 6. Ablation study of two factors in HalluciDoctor.

5.6. In-depth Analysis

Robustness Analysis of HalluciDoctor. Additionally, we investigated the robustness of HalluciDoctor when applied to other machine-generated datasets, such as MiniGPT4-Instruction [44]. As illustrated in Table 4, HalluciDoctor similarly reduces hallucination frequency and enhances model performance, demonstrating our method's robustness across various machine-generated visual instructions.

Analysis of Instruction Expansion Factors. To dissect the influence of various factors in our counterfactual instruction expansion, we incrementally removed them and presented the associated results in Figure 6. The absence of enhancement factors for balancing tail co-occurrences leads MLLMs to more hallucinations, while the lack of inhibitory factors leads to excessive unreasonable instructions, diminishing MLLM performance in contextual understanding.

Visualization Results. In Figure 4, we present a case where MLLM progressively enhances response quality after HalluciDoctor mitigates the hallucinatory toxicity from visual instruction data. In this case, HalluciDoctor effectively reduces hallucinatory toxicity introduced by LLaVA (*e.g.*, pedestrian crossing the street). Furthermore, with the aid of the more robust LLaVA++, MLLMs reduce the impact of specious correlations and enhance the perception of fine-grained (*e.g.*, shop's name) and unusual content (*e.g.*, cow out front of shop). More case analyses are in Appendix D.

6. Conclusions

In this paper, we initially delve deep into the underlying hallucination phenomena in machine-generated visual instruction data. We introduce a flexible framework, HalluciDoctor, that exploits a question-based cross-checking paradigm to detect and eliminate potential hallucinations automatically. Additionally, we pinpoint the co-occurrence issue leading to hallucinations and augment the MLLMs' resistance to such errors through the expansion of counterfactual instruction. The extensive experimental results by both automatic metrics and human evaluations demonstrate the superiority of our approach in dispelling various hallucinations and retaining MLLM's open-ended capabilities.

Acknowledgment. This work was supported by the National Natural Science Foundation of China (U2336212), Key Research and Development Projects in Zhejiang Province (No. 2024C01106), the National Key Research and Development Project of China (2018AAA0101900), and Research funding from FinVolution Group. We thank all the reviewers for their valuable comments.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 1

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 2

[3] Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Boerschinger, and Tal Schuster. Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. *arXiv preprint arXiv:2202.07654*, 2022. 5

[4] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023. 1

[5] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 1, 2, 5

[6] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 7

[7] Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. *arXiv preprint arXiv:2308.06394*, 2023. 2, 3

[8] Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven Hoi. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10867–10877, 2023. 2

[9] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 6

[10] Tiep Le, Vasudev Lal, and Phillip Howard. Coco-counterfactuals: Automatically constructed counterfactual examples for image-text pairs. *arXiv preprint arXiv:2309.14356*, 2023. 2

[11] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023. 2

[12] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 1, 3

[13] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1, 2, 5

[14] Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Hanwang Zhang, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, and Yueting Zhuang. Fine-tuning multimodal llms to follow zero-shot demonstrative instructions. In *The Twelfth International Conference on Learning Representations*, 2023. 1, 2

[15] 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. 2, 3, 6

[16] Yongqi Li, Mayi Xu, Xin Miao, Shen Zhou, and Tieyun Qian. Large language models as counterfactual generator: Strengths and weaknesses. *arXiv preprint arXiv:2305.14791*, 2023. 2

[17] Zhuang Li, Yuyang Chai, Terry Zhuo Yue, Lizhen Qu, Gholamreza Haffari, Fei Li, Donghong Ji, and Quan Hung Tran. Factual: A benchmark for faithful and consistent textual scene graph parsing. *arXiv preprint arXiv:2305.17497*, 2023. 4

[18] Ning Liao, Shaofeng Zhang, Renqiu Xia, Bo Zhang, Min Cao, Yu Qiao, and Junchi Yan. Revo-lion: Evaluating and refining vision-language instruction tuning datasets. *arXiv preprint arXiv:2310.06594*, 2023. 2, 3

[19] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[20] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023. 2, 3, 6, 7

[21] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 1, 2, 3, 6, 7

[22] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 3

[23] Jiaying Lu, Jinmeng Rao, Kezhen Chen, Xiaoyuan Guo, Yawen Zhang, Baochen Sun, Carl Yang, and Jie Yang. Evaluation and mitigation of agnosia in multimodal large language models. *arXiv preprint arXiv:2309.04041*, 2023. 2

[24] OpenAI. ChatGPT. https://openai.com/blog/chatgpt/, 2023. 3

[25] OpenAI. GPT-4. https://openai.com/gpt-4, 2023. 2, 3

[26] 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. 2

[27] Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. *Advances in neural information processing systems*, 33:857–869, 2020. 5

[28] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023. 1

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

[30] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 1

[31] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021. 1, 2

[32] Bin Wang, Fan Wu, Xiao Han, Jiahui Peng, Huaping Zhong, Pan Zhang, Xiaoyi Dong, Weijia Li, Wei Li, Jiaqi Wang, et al. Vigc: Visual instruction generation and correction. *arXiv preprint arXiv:2308.12714*, 2023. 2, 3, 6, 7

[33] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022. 2

[34] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022. 2

[35] Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. *arXiv preprint arXiv:2307.02477*, 2023. 2

[36] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 1, 2, 6, 7

[37] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023. 2

[38] Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*, 2023. 2

[39] Qifan Yu, Juncheng Li, Yu Wu, Siliang Tang, Wei Ji, and Yueting Zhuang. Visually-prompted language model for fine-grained scene graph generation in an open world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21560–21571, 2023. 2

[40] An Zhang, Wenchang Ma, Xiang Wang, and Tat-Seng Chua. Incorporating bias-aware margins into contrastive loss for collaborative filtering. *Advances in Neural Information Processing Systems*, 35:7866–7878, 2022. 7

[41] Letian Zhang, Xiaotong Zhai, Zhongkai Zhao, Xin Wen, and Bingchen Zhao. What if the tv was off? examining counterfactual reasoning abilities of multi-modal language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4629–4633, 2023. 2

[42] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023. 2

[43] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*, 2023. 3, 6, 7

[44] 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, 3, 5, 6, 8