HALLUSIONBENCH: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in Large Vision-Language Models

Tianrui Guan* Fuxiao Liu* Xiyang Wu Ruiqi Xian Zongxia Li Xiaoyu Liu Xijun Wang
Lichang Chen Furong Huang Yaser Yacoob Dinesh Manocha Tianyi Zhou

University of Maryland, College Park
{rayguan, fl3es, wuxiyang, rxian, zli12321, xliu1231, xijun
bobchen, furongh, yaser, dmanocha, tianyi}@umd.edu

Abstract

*We introduce "*HALLUSIONBENCH[1]*," a comprehensive benchmark designed for the evaluation of image-context reasoning. This benchmark presents significant challenges to advanced large visual-language models (LVLMs), such as GPT-4V(ision), Gemini Pro Vision, Claude 3, and LLaVA-1.5, by emphasizing nuanced understanding and interpretation of visual data. The benchmark comprises 346 images paired with 1129 questions, all meticulously crafted by human experts. We introduce a novel structure for these visual questions designed to establish control groups. This structure enables us to conduct a quantitative analysis of the models' response tendencies, logical consistency, and various failure modes. In our evaluation on* HALLUSIONBENCH*, we benchmarked 15 different models, highlighting a 31.42% question-pair accuracy achieved by the state-of-the-art GPT-4V. Notably, all other evaluated models achieve accuracy below 16%. Moreover, our analysis not only highlights the observed failure modes, including language hallucination and visual illusion but also deepens an understanding of these pitfalls. Our comprehensive case studies within* HALLUSIONBENCH *shed light on the challenges of hallucination and illusion in LVLMs. Based on these insights, we suggest potential pathways for their future improvement. The benchmark and codebase can be accessed at https://github.com/tianyi-lab/HallusionBench.*

1. Introduction

In recent years, Large Language Models (LLMs) [9, 10, 26, 40, 45, 46, 61] have revolutionized the field of machine learning with the ability of language understanding and content generation, offering unprecedented capabilities and potentials across a multitude of applications. The integration of LLMs with computer vision systems has given rise to Large Vision-Language Models (LVLMs) [6, 8, 22, 27, 28, 33, 40, 41, 49, 50, 55, 63]. These models have demonstrated profound capabilities in various applications and significantly enhance the performance in image reasoning tasks [5, 18, 20, 30, 31, 36, 38, 42, 47]. However, the hallucination issue of LLMs [58] is regarded as a challenging and unsolved problem, which leads to many issues when we integrate LLMs with vision techniques.

While LVLMs like GPT-4V(ision) [48] and LLaVA-1.5 [32] excel in various applications, they are hindered by a pronounced language bias. This bias stems from instances where knowledge priors conflict with the visual context [24, 29, 57]. Similarly, models such as LLaVA-1.5 [32] and mPLUG-Owl [50] are prone to giving affirmative answers regardless of the actual content of questions [24]. The distinct failure modes of different VLMs highlight the need for specific improvements. Recognizing and understanding these limitations and failure types is imperative for advancing these models and striking a delicate balance between knowledge priors and contextual understanding.

When exploring those LVLMs, we observe that their strong language bias often overshadows visual information, leading to an overreliance on language priors rather than the visual context. To study this phenomenon, we use the term "***Language Hallucination***," which refers to conclusions drawn without visual input. On the other hand, the vision components within the limited ability in LVLMs can give rise to "***Visual Illusion***", where visual inputs can be misinterpreted, leading to overconfident yet erroneous assertions by the model.

Main Contributions: Recognizing the need to comprehend why an LVLM fails and address these issues, we present HALLUSIONBENCH, a carefully crafted benchmark designed to explore the complexities of image-context rea-

*Equal contribution.
[1]"Hallusion" is a portmanteau of "**hall**ucination" and "i**llusion**."

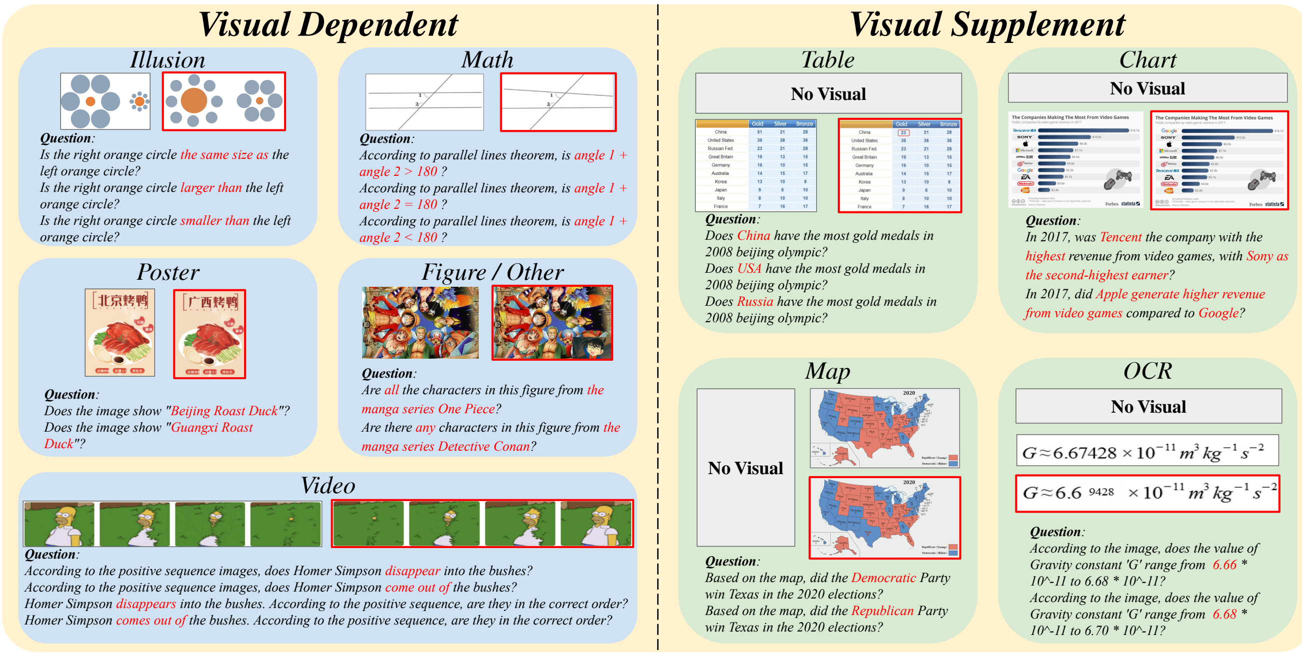


Figure 1. Data samples of HALLUSIONBENCH, which contains diverse topics, visual modalities. Human-edited images are in RED, resulting in different correct answers to the questions.

soning in depth and expose various problems with respect to current LVLMs, as shown in Fig. 1. Our design of the visual-question (VQ) pairs, unique in format, facilitates a quantitative analysis of the models' failures, enabling a more thorough evaluation. This investigation sheds light on existing limitations and lays the groundwork for future improvements, aiming to make the next generation of LVLMs more robust, balanced, and precise. The novelties of our work include:

1. We introduce HALLUSIONBENCH, the first advanced diagnostic suite tailored to systematically dissect and analyze the diverse failure modes of LVLMs. HALLUSIONBENCH consists of approximately 1129 handcrafted visual question-answer (VQA) pairs, featuring 165 original images and 181 images expertly modified by human professionals. Moving beyond the traditional metrics of correctness and accuracy, our VQA pairs are thoughtfully formulated with an innovative structure. This approach enables us to quantitatively analyze specific dimensions and aspects where current models falter.
2. We evaluate 15 most recent methods on HALLUSIONBENCH. Our benchmark presents formidable challenges to existing methods. Notably, the SoTA GPT-4V achieves merely a 31.42% Question Pair Accuracy, while the performance of all other methods falls below 16%.
3. We explore HALLUSIONBENCH and provide an in-depth analysis of examples on which the SoTA LVLMs, such as GPT-4V and LLaVA-1.5 fail. We also provide insights on different issues that existing LVLMs are facing based on the quantitative analysis enabled by HALLUSIONBENCH. In our exploration of HALLUSIONBENCH, we conduct a detailed analysis of instances where SoTA LVLMs, including GPT-4V and LLaVA-1.5, fall short. Additionally, our investigation leverages the quantitative capabilities of HALLUSIONBENCH to shed light on various issues currently challenging existing LVLMs.

2. Related Work

2.1. Large Multi-Modal Models

Large Language Models have been a major advancement, leading to new ways to understand not just text but other things like images, all in one large system. For example, Flamingo [3] has many capabilities, combining a vision part that doesn't change with a big language model that has a special feature for understanding both images and words together. Another model, PaLM-E [13], mixes visual information directly into the already powerful PaLM model, which has 520 billion parameters, making it effective in real-world uses. Most recently, researchers have been creating high-quality, diverse multi-modal datasets from GPT4 and GPT-4V [48] to fine-tune open-source LVLMs, including LLaVA [33], MiniGPT4 [63], Mplug-Owl [50], LRV-Instruction [29], LLaVAR [60] and other works [12, 25, 37, 52].

2.2. Hallucination in LVLMs

Hallucination typically refers to situations where the generated responses contain information that is not present in the visual content. Prior research primarily examines two areas: detecting and evaluating hallucinations [24, 58, 59], and methods to reduce them [29, 43, 53]. Early methods include training classifiers to identify hallucinations or comparing output with accurate answers to detect inaccuracies. To mitigate hallucinations, efforts have been made to improve data gathering and training procedures. For example, LRV-Instruction [29] creates balanced positive and negative instructions to finetune LVLMs. VIGC [43] uses an iterative process to generate concise answers and combine them, aiming for detailed yet accurate responses. Similarly, Woodpecker [53] introduces a training-free method to pick out and correct hallucinations from the generated text.

2.3. Benchmarks for Large VL Models

Traditional Visual Language (VL) benchmarks are designed to assess distinct skills, including visual recognition [17], image description [2, 28], and so on. However, with the advent of advanced LVLMs, traditional evaluation metrics often fall short of providing a detailed ability assessment. This problem is further exacerbated by their inability to match the given answer accurately, leading to significant robustness issues. To address these challenges, research communities have introduced a series of benchmarks, including MME [15], MMBench [34], MM-Vet [54], SEED-Bench [21], GAVIE [29], and LAMM-Bench [14]. These benchmarks systematically structure and evaluate complex multi-modal tasks. Different from POPE [24] and GAVIE [29] evaluating the object hallucinations of LVLMs, HALLUSIONBENCH is the first human-annotated analytical benchmark focusing on diagnosing both the visual illusion and knowledge hallucination of LVLMs.

3. HALLUSIONBENCH Construction

We present HALLUSIONBENCH, the first benchmark designed to examine visual illusion and knowledge hallucination of LVLMs and analyze the potential failure modes based on each hand-crafted example pair. HALLUSIONBENCH consists of 455 visual-question control pairs, including 346 different figures and a total of 1129 questions on diverse topics (including ***food, math, geometry, statistics, geography, sports, cartoon, famous illusions, movie, meme,*** *etc.*) and formats (including ***logo, poster, figure, charts, table, map, consecutive images,*** *etc.*). In the following sections, we first provide the guidelines for dataset construction based on different visual question types. Second, we will describe the data and annotation structure of HALLUSIONBENCH. Finally, we will describe the statistics of our dataset.

3.1. Visual Question Taxonomy

Our aim is to develop a multimodal image-context reasoning benchmark to investigate the potent language bias inherent in LVLMs, which can sometimes overshadow the visual context. We define the two categories of visual questions: ***Visual Dependent*** and ***Visual Supplement***.

3.1.1 Visual Dependent Questions

The ***Visual Dependent*** questions are defined as ***questions that do not have an affirmative answer without the visual context.*** Such questions ask about the image itself or something within the image. For example, there is no clear answer to *"Is the right orange circle the same size as the left orange circle?"* without an image to provide more context.
Guideline: Under this setting, our benchmark is designed to evaluate visual commonsense knowledge and visual reasoning skills. Our exploration and dataset construction are guided by the following questions:

1. *How good are the visual understanding and reasoning skills of the model?*
2. *How does the parametric memory of the model affect its response to a question?*
3. *Is the model able to capture the temporal relation of multiple images?*

3.1.2 Visual Supplement Questions

The *Visual Supplement* questions are ***questions that can be answered without the visual input; the visual component merely provides supplemental information or corrections.*** For example, some LVLMs can answer *"Is New Mexico state larger than Texas state?"* using the prior knowledge in their parametric memory without a map of the US.
***Guideline*:** Under this setting, our benchmark is designed to evaluate visual reasoning ability and the balance between parametric memory and image context. Our exploration and dataset construction under this category is guided by the following questions:

1. *When the model lacks the prior knowledge or answer in the parametric memory of its language module, does the model (still) hallucinate about the images?*
2. *When the model's language module has sufficient prior knowledge in its parametric memory or directly knows the answer, does it still enhance its response by gathering extra information from the visual supplement (especially when the prior knowledge conflicts with the visual input or the parametric memory is outdated)?*
3. *How well can the model interpret a visual input with dense information (i.e., a graph, chart, map, etc.) for question answering? What types of image manipulation might impede or distort visual information extraction?*

3.2. Visual, Question, and Annotation Structures

Notations: Let $(I, q) \in \mathcal{V} \subseteq \mathbb{I} \times \mathbb{Q}$ be the tuple of the image $I \in \mathbb{I}$ and question $q \in \mathbb{Q}$, where $\mathcal{V}$ is the set of valid VQ pairs. Let N be the number of original images obtained from the Internet, and $\mathbb{I}_o = \{I_{(i,0)}\}_{0<i\leq N}$ be the set of those

		No Visual	Original Visual	Edited Visual	Overall	
Visual Dependent	*Illusion*	-	72	72	144	591
	Math	-	54	54	108	
	Video	-	69	101	170	
	Poster	-	43	46	89	
	Others	-	39	41	80	
Visual Supplement	*Chart*	76	68	62	206	538
	Table	43	43	69	155	
	Map	32	32	32	96	
	OCR	27	27	27	81	
Overall		178	447	504	1129	

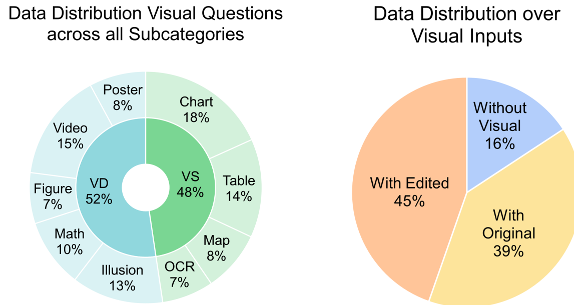


Figure 2. **Statistics of HALLUSIONBENCH:** We show the number of questions in the table (*left*), and the distribution of visual questions across each subcategory of Visual Dependent (VD) and Visual Supplement (VS) (*middle*) and visual input types categorized by no visual, original, and edited images (*right*). HALLUSIONBENCH covers a diverse visual format and nearly half of the images are manually edited.

Benchmarks	Visaul Format	# Total QA	# H-Edited QA	# Total Img.	# H-Edited Img.	Control Pair?	Purpose
Lynx-Bench [56]	Image,Video	450	450	450	0	✗	Image&Video QA Evaluation
SciGraphQA [23]	Image	295K	0	657K	0	✗	Scientific Chart QA Evaluation
MathVista [35]	Image	6141	0	5487	0	✗	Math Reasoning Evaluation
MME [15]	Image	1457	1457	1187	0	✗	Comprehensive Evaluation
POPE [24]	Image	3000	0	500	0	✗	Object Hallucination
M-HalDetect [19]	Image	4000	0	4000	0	✗	Object Hallucination
GAVIE [29]	Image	1000	0	1000	0	✗	Object Hallucination
Bingo [11]	Image	370	370	308	N/A	✓	Hallucination, Bias
HALLUSIONBENCH	Image, Video **Image Pairs**	1129	1129	346	**181**	✓	**Visual Illusion, Language Hallucination, Quantitative Analysis and Diagnosis**

Table 1. **Comparison of HALLUSIONBENCH with most recent VL benchmarks:** HALLUSIONBENCH is the **first** and the **only** benchmark that focuses on control-group analysis by carefully editing each image in the database manually. "# H-Edited QA" means Human-edited question-answer pairs. "# H-Edited Img" means Human-edited images. *N/A* denotes that the information is not provided.

original images. We define $\mathbb{I}'_i = \{I_{(i,j)}\}_{0<j\leq N_i}$ be the set of images modified from $I_{(i,0)}$, and I_0 be an empty image. The entire images set $\mathbb{I} = \{I_0\} \bigcup \mathbb{I}_o \bigcup (\bigcup_{0<i\leq N} \mathbb{I}'_i)$.

Let $\mathbb{Q}_i = \{q_{(i,k)}\}_{0<k\leq M_i}$ be the set of questions that can be applied to any image in $\mathbb{I}_i$, which is defined differently for Visual Dependent (*VD*) and Visual Supplement (*VS*):

$$\mathbb{I}_i = \begin{cases} \{I_{(i,0)}\} \bigcup \mathbb{I}'_i & \text{for } VD \\ \{I_0, I_{(i,0)}\} \bigcup \mathbb{I}'_i & \text{for } VS \end{cases} \tag{1}$$

To facilitate evaluation, all questions are formulated as Yes/No questions (Fig. 1). We annotate each visual-question with a binary answer $y(I, q) \in \{\text{“yes”}, \text{“no”}\}$.

3.3. Dataset Statistics

Following the annotation structure and guidelines above, we ask human experts to collect 346 images with diverse topics and types manually. As shown Fig. 2, *Visual Dependent* has 591 questions, including *videos, illusion, math, posters, logos, cartoons*, and *others*; *Visual Supplement* has 538 questions, including *charts, tables, maps*, and *OCR*. Furthermore, Fig. 2 (*right*) describes the distribution of the questions without visual input (16%), with original online images (39%), and with visual input edited by human experts (45%). Our image manipulation strategies contain *image flipping, order reversing, masking, optical character editing, object editing*, and *color editing*. Additionally, each image has 3.26 questions on average. Fig. 2 (*left*) provides more details on the number of questions in each topic and visual input category.

3.4. Uniqueness of HALLUSIONBENCH

The main comparison between HALLUSIONBENCH and existing benchmarks is presented in Tab. 1. As it shows, there is a notable gap between existing benchmarks[11, 19, 24, 29] and HALLUSIONBENCH in hallucination evaluation, as existing benchmarks primarily focus on object hallucinations, limited topics, and visual input types. Our dataset, HALLUSIONBENCH, is therefore motivated to bridge this gap by providing more topics, more image types, and more visual input modalities, including both images and videos. Additionally, our human experts carefully select each image and write question-answer pairs. We are also the first work to include human-edited images to assess the robustness of current LVLMs. Additionally, unlike existing benchmarks, HALLUSIONBENCH focuses on evaluating both language hallucinations and visual illusions, moving beyond the narrow scope of object hallucinations [19, 24, 29].

4. HALLUSIONBENCH Evaluation Suite

4.1. Text-Only GPT4-Assisted Evaluation

Notations: Let $\mathcal{M}(I, q) \in \{\text{“yes”}, \text{“no”}, \text{“uncertain”}\}$ be the parsed output answer by a VLM $\mathcal{M}$ for an image-question pair (I, q). GPT-4 $GPT(\mathcal{M}(I, q),\ y(I, q))$ then judges the answer $\mathcal{M}(I, q)$ based on the ground truth $y(I, q) \in \{\text{“yes”}, \text{“no”}\}$ and outputs *Incorrect (0)*, *Correct (1)*, or *Uncertain (2)* if the predicted response is ambiguous.

The prompt for the GPT-4 judge is designed as:

Imagine you are an intelligent teacher. Thoroughly read the question, reference answer, and the prediction answer to ensure a clear understanding of the information provided. Assess the correctness of the predictions. If the prediction answer does not conflict with the reference answer, please generate "correct". If the prediction answer conflicts with the reference answer, please generate "incorrect". If the prediction answer is unclear about the answer, please generate "unclear".

For each sample, we fill the template with its question, ground truth, and LVLM output. By taking the filled prompt into GPT-4, GPT-4 will generate "correct", "incorrect" or "unclear" for the sample. It is found that outputs of GPT-4 still exist variance, although the temperature is set as 0. Therefore, we utilize GPT-4 to evaluate the outputs of LLMs 3 times and report average scores.

Comparison with Human Evaluation: To demonstrate that our GPT4-Assisted evaluation is effective, we obtain the responses from GPT-4V [48] and LLaVA-1.5 [32], and manually evaluate the correctness of their responses. We label the responses with *Incorrect (0)*, *Correct (1)*, and *Uncertain (2)* if the answer is ambiguous. As shown in the first two rows of Tab. 2 and Tab. 3, the negligible difference proves that the GPT4-assisted method aligns well with human judgment.

4.2. Correctness Evaluation Metrics

Since the focus of our benchmark is on hallucination and illusion, not the span of knowledge, we consider an *uncertain* answer acceptable when there is no visual input under the *Visual Supplement* category. For the final accuracy score, we convert the correctness into a binary value $b_{\mathcal{M}} \in \{0, 1\}$:

$$b_{\mathcal{M}}(I,q) = \begin{cases} GPT(\mathcal{M}(I,q),\ y(I,q)) & \text{if } GPT(\mathcal{M},\ y) \leq 1 \\ 1 & \text{else if } I = I_0 \\ 0 & \text{otherwise} \end{cases}, \tag{2}$$

Let $(I, q) \in \mathcal{V} \subseteq \mathbb{I} \times \mathbb{Q}$ be the tuple of the image $I \in \mathbb{I}$ and question $q \in \mathbb{Q}$, where $\mathcal{V}$ is the set of valid visual-question pairs. Let $\mathbb{1}(\cdot)$ be the indicator function.

All accuracy:

$$aAcc = \frac{\sum_{(I,q)\in\mathcal{V}} b_{\mathcal{M}}(I,q)}{|\mathcal{V}|} \tag{3}$$

Figure Accuracy:

$$fAcc = \frac{\sum_{i,j} \mathbb{1}(\bigwedge_{q\in\mathbb{Q}_i} b_{\mathcal{M}}(I_{(i,j)}, q))}{|\mathbb{I}|} \tag{4}$$

Question Pair Accuracy:

$$qAcc = \frac{\sum_{i,k} \mathbb{1}(\bigwedge_{I\in\mathbb{I}_i} b_{\mathcal{M}}(I, q_{(i,k)}))}{|\mathbb{Q}|} \tag{5}$$

4.3. Analytical Evaluation Criteria

In addition to the accuracy metrics, we introduce three analytical criteria to measure and diagnose the failures of LVLMs, *Yes/No Bias Test*, *Consistency Test*, and *Diagnostic Test*. Instead of examining and analyzing each failed case qualitatively, we propose these novel quantitative measurements through the unique design of our question sets. These tests are listed in the order of complexity, so the latter test would not be as useful and insightful if the former basic test failed.

4.3.1 Yes / No Bias Test

According to [24], some models [16, 32, 50] tend to respond with "yes" in most cases. No further analysis is necessary if the model has a very strong bias or tendency to answer one way regardless of the actual question, so we design two criteria to reveal such preference of the model.

Yes Percentage Difference (Pct. Diff) $d_y \in [-1, 1]$**:**

$$d_y = \frac{\sum\limits_{(I,q)\in\mathcal{V}} \left[\mathbb{1}\big(\mathcal{M}(I,q) = \text{“yes”}\big) - \mathbb{1}\big(y(I,q) = \text{“yes”}\big)\right]}{|\mathcal{V}|}, \tag{6}$$

d_y represents the difference between the predicted and actual number of "Yes" in the question set. The model is more biased when $|d_y|$ is close to 1.

False Positive Ratio (FP Ratio) $r_{fp} \in [0, 1]$**:**

$$r_{fp} = \frac{\sum_{(I,q)\in\mathcal{W}} \mathbb{1}\big(\mathcal{M}(I,q) = \text{“yes"}\big)}{|\mathcal{W}|}, \tag{7}$$

where $\mathcal{W} = \{(I, q) \in \mathcal{V} \mid b_{\mathcal{M}}(I, q) = 0\}$ is the set of incorrect visual questions. r_{fp} measures how likely the model responses with "Yes" out of all incorrect responses. The model is more robust when r_{fp} is close to 0.5.

4.3.2 Consistency Test

The goal of the consistency test is to test the logical consistency of responses and make sure questions are not answered based on random guesses. Many questions $\mathbb{Q}^i$ from root $\mathcal{R}^i$ are logically consistent: for example, "Is the left segment longer than/shorter than/equal to the right segment?" The consistency test is implemented and measured using *fAcc* (Metrics 4). We design the question set $\mathbb{Q}_i$ to be logically correlated over a figure. Therefore, we consider the model *inconsistent* when only some of the questions in $\mathbb{Q}_i$ are correct. In other cases, the model would be consistently correct or consistently wrong.

4.3.3 Language Hallucination and Visual Illusion

Before we dive into the diagnostic test, we categorize the failures into two major types based on the failed cases:

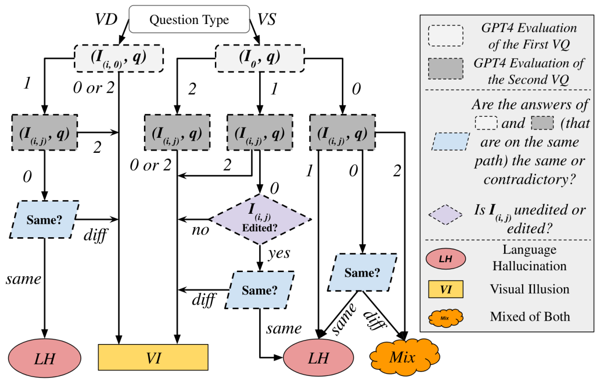


Figure 3. **Decision Tree to Diagnose Failure Types:** Based on the correctness of two questions in a control pair, and the difference of their responses, we use this decision tree to analyze the failure. The output of *GPT4 Evalution* could be *Incorrect (0)*, *Correct (1)*, or *Uncertain (2)* if the predicted response is ambiguous.

Language Hallucination *refers to perceptions formed without relevant visual input.* In language hallucination, the model makes false prior assumptions about the input and image context based on its parametric memory. The model should respond based on how the question is framed instead of ignoring it or making false assumptions about the image.

Visual Illusion *denotes the misinterpretation of accurate visual information.* Visual illusion comes from the failure to recognize and understand the input image visually. The model could not obtain accurate information or reason about the image correctly.

4.3.4 Diagnostic Test

To study the issue of language hallucination and language illusion, we analyze the responses and correctness of both visual questions within a *VQ Control Pairs* and divide incorrect responses into three categories: *Language Hallucination*, *Visual Illusion*, and *Mixed / Uncertain*. We measure the percentage of those failures out of all failed cases.

Control Pair: The control pair will always contain an original image for *visual dependent* questions or an empty image (no visual) for *visual supplement* questions. The other question in the control pair may have an edited image (or an original image for *VS* question). The response to this question would provide more information on whether the answer exists in the parametric knowledge or if the model has seen it in the training data. In addition, we can examine whether the response remains the same after editing the original image to obtain more insights into the failures, which is more informative than checking a single visual question alone. In Fig. 3, we provide a decision tree to determine the type of failure for a control pair. We consider the following principles when assigning the failure types:

1. For *visual dependent (VD)* questions, or *visual supplement (VS)* questions that have visual inputs, if the response is incorrect or uncertain, the failure could be ***visual illusion***, since the model could not extract from the visual information correctly.
2. For *visual supplement (VS)* questions that don't have visual inputs, if the response gives a certain but wrong answer, we attribute it to ***language hallucination***.
3. If the model responds to the original image (or no image) correctly and has the same response to the edited image (which is contrary to common sense), it means that the parametric knowledge overtakes the actual image input. Therefore, we also attribute the failure to ***language hallucination***.

We will include some examples in the supplemental material.

5. Experimental Results

5.1. Models

We conduct massive experiments on HALLUSIONBENCH to evaluate a total of 15 LVLMs, including GPT-4V [1], LLaVA-1.5 [32], Gemini Pro Vision [39], Claude 3 [4], MiniGPT4 [63], MiniGPT5 [62], GiT [44], InstructBLIP [12], Qwen-VL [7], mPLUG-Owl-v1 [50], mPLUG-Owl-v2 [51], LRV-Instruction [29], BLIP2 [22], BLIP2-T5 [22], and Open-Flamingo [3]. We also include *Random Chance* (i.e. randomly choose *Yes* or *No*) as a baseline.

5.2. Result Analysis

We compare the performance of several models, including both closed-source models and open-sourced models. Results are given in Tab. 2, Tab. 3 and Fig. 4. Additionally, we established a human expert evaluation to assess the effectiveness of text-only GPT4-assisted evaluation.

Correctness Evaluation. As shown in Tab. 2, GPT-4V outperforms all the open-sourced LVLMs by a large margin except the *Hard Accuracy*. *Hard Accuracy* measures the models' ability to understand human-edited images from HALLUSIONBENCH. The poor accuracy demonstrates the challenges of our image manipulations for GPT-4V and other open-source LVLMs. In the open-sourced models, we investigate if expanding the size (0.8B to 13B) of the LLM backbone can mitigate object existence hallucination. As detailed in Tab. 2, there is a noticeable reduction in hallucination as the model size increases, like LLaVA-1.5 and BLIP2-T5. Among models with a size of less than 10B, InstructBLIP and mPLUG-Owl-v2 are the best-performing ones. InstructBLIP, leveraging the BLIP-2 architecture and enhanced through instruction fine-tuning across 26 diverse datasets, demonstrates that a broader and more extensive training set can substantially enhance performance. The boosting performance of mPLUG-Owl-v2 compared with mPLUG-Owl-v1 can be attributed to its novel module, which utilizes the language decoder acting as a universal interface for managing different modalities.

Method	# Parameter	Evaluation	Question Pair Accuracy (*qAcc*) ↑	Figure Accuracy (*fAcc*) ↑	Easy Accuracy (Easy *aAcc*) ↑	Hard Accuracy (Hard *aAcc*) ↑	All Accuracy (*aAcc*) ↑
GPT4V [1] (Oct 2023)	-	Human	31.42	44.22	79.56	38.37	67.58
		GPT4-Assisted	**28.79**	**39.88**	**75.60**	37.67	**65.28**
LLaVA-1.5 [32]	13B	Human	9.45	25.43	50.77	29.07	47.12
		GPT4-Assisted	10.55	**24.86**	**49.67**	29.77	46.94
Claude 3 [4]	-	GPT4-Assisted	**21.76**	**28.61**	**55.16**	**41.40**	**56.86**
Gemini Pro Vision [39] (Dec 2023)	-	GPT4-Assisted	7.69	8.67	35.60	30.23	36.85
BLIP2-T5 [22]	12.1B	GPT4-Assisted	**15.16**	20.52	45.49	**43.49**	**48.09**
Qwen-VL [7]	9.6B	GPT4-Assisted	5.93	6.65	31.43	24.88	39.15
Open-Flamingo [3]	9B	GPT4-Assisted	6.37	11.27	39.56	27.21	38.44
MiniGPT5 [62]	8.2B	GPT4-Assisted	10.55	9.83	36.04	28.37	40.30
MiniGPT4 [63]	8.2B	GPT4-Assisted	8.79	10.12	31.87	27.67	35.78
InstructBLIP [12]	8.2B	GPT4-Assisted	9.45	10.11	35.60	**45.12**	45.26
BLIP2 [22]	8.2B	GPT4-Assisted	5.05	12.43	33.85	40.70	40.48
mPLUG_Owl-v2 [51]	8.2B	GPT4-Assisted	13.85	19.94	44.84	39.07	47.30
mPLUG_Owl-v1 [50]	7.2B	GPT4-Assisted	9.45	10.40	39.34	29.77	43.93
LRV_Instruction [29]	7.2B	GPT4-Assisted	8.79	13.01	39.78	27.44	42.78
GIT [44]	0.8B	GPT4-Assisted	5.27	6.36	26.81	31.86	34.37
Random Chance	-	GPT4-Assisted	15.60	18.21	39.12	39.06	45.96

Table 2. **Correctness Leaderboard on HALLUSIONBENCH with various LVLMs:** All the numbers are presented in % and the full score is 100%. Hard questions refer to the edited images. We highlight the Top 3 models with the GPT4-assisted evaluation.

			Yes/No Bias		Consistency			Language and Vision Diagnosis		
Method	# Parameter	Evaluation	Pct. Diff (∼ 0)	FP Ratio (∼ 0.5)	Correct ↑	Inconsistent ↓	Wrong ↑	Language Hallucination	Visual Illusion	Mixed
GPT4V [1] (Oct 2023)	-	Human	0.066	0.60	44.22	32.66	23.12	21.86	46.17	31.97
		GPT4-Assisted	**0.058**	0.58	**39.88**	**38.15**	**21.97**	22.19	45.66	32.14
LLaVA-1.5 [32]	13B	Human	0.27	0.76	25.43	42.49	32.08	25.63	51.42	22.95
		GPT4-Assisted	0.26	0.75	**24.86**	**45.38**	29.77	26.71	51.09	22.20
Claude 3 [4]	-	GPT4-Assisted	0.063	**0.57**	**28.61**	**49.42**	**21.97**	19.10	59.14	21.77
Gemini Pro Vision [39] (Dec 2023)	-	GPT4-Assisted	**-0.02**	**0.48**	8.67	56.94	34.39	25.95	49.37	24.68
BLIP2-T5 [22]	12.1B	GPT4-Assisted	0.08	0.58	20.52	59.54	19.94	41.64	40.44	17.92
Qwen-VL [7]	9.6B	GPT4-Assisted	0.12	0.60	6.65	50.29	43.06	0.87	88.06	11.06
Open-Flamingo [3]	9B	GPT4-Assisted	0.33	0.77	11.27	59.83	28.90	30.07	48.06	21.87
MiniGPT5 [62]	8.2B	GPT4-Assisted	0.28	0.71	9.83	56.36	33.82	10.09	73.44	16.47
MiniGPT4 [63]	8.2B	GPT4-Assisted	0.19	0.65	10.12	57.80	32.08	23.59	56.55	19.86
InstructBLIP [12]	8.2B	GPT4-Assisted	-0.13	0.38	10.12	68.50	**21.39**	29.29	54.53	16.18
BLIP2 [22]	8.2B	GPT4-Assisted	0.18	0.65	12.43	63.01	24.57	39.14	43.45	17.41
mPLUG_Owl-v2 [51]	8.2B	GPT4-Assisted	0.25	0.77	19.94	58.09	21.97	28.24	50.42	21.34
mPLUG_Owl-v1 [50]	7.2B	GPT4-Assisted	0.32	0.79	10.40	60.12	29.48	3.95	78.36	17.69
LRV_Instruction [29]	7.2B	GPT4-Assisted	0.26	0.73	13.01	53.47	33.53	4.49	76.47	19.04
GIT [44]	0.8B	GPT4-Assisted	**0.04**	**0.53**	6.36	53.76	39.88	30.90	58.30	10.80
Random Chance	-	GPT4-Assisted	0.08	0.57	18.20	57.51	24.28	-	-	-

Table 3. **Analytical Evaluation Results on HALLUSIONBENCH with various LVLMs:** *Pct. Diff* ranges from [-1, 1]. The model is more biased when *Pct. Diff* is close to -1 or 1. *FP Ratio* ranges from [0, 1]. The model is more robust when *FP Ratio* is close to 0.5. All the other metrics are presented in %, and the full score is 100%. We highlight the Top 3 models with the GPT4-assisted evaluation.

Yes/No Bias. Another observation is that GPT-4V, BLIP2-T5, and mPLUG-Owl-v2 outperform *Random Choice* in both question pair accuracy, figure pair accuracy, and question level accuracy. Other models, such as Qwen-VL and MiniGPT4, perform even worse than *Random Choice*. This indicates their visual reasoning abilities are still limited. However, LLaVA-1.5 outperforms *Random Choice* while achieving poor results in both question pair accuracy and figure pair accuracy. We attribute this phenomenon to the fact that LLaVA-1.5 tends to answer *Yes*. This assumption is supported by the low *Yes Percentage Difference* and *False Positive Ratio* of LLaVA-1.5 in *Yes/No Bias Test* from Tab. 3. Besides, we find that Open-Flamingo and mPLUG-Owl-v1 also tend to answer *Yes* with the high *Yes Percentage Difference* and *False Positive Ratio*. Inspired by [29], one possible reason is that these LVLMs lack balanced positive and negative instructions in their training set. We also attribute the poor performance of these LVLMs to the scarcity of human-edited images in their training set since most LVLMs only utilize original images from existing datasets.

Language and Vision Diagnosis. We report fine-grained scores of six prominent LVLMs across different visual inputs in Fig. 4. Results show that *Math*, *Illusion*, and *Video* is the most challenging format for current LVLMs, including GPT-4V. From Fig. 5 (top), we found both GPT-4V and LLaVA-1.5 are unable to correctly recognize regular triangles, meaning that geometry and math are still a challenging task for GPT-4V. From Fig. 5 (middle), we found GPT-4V is

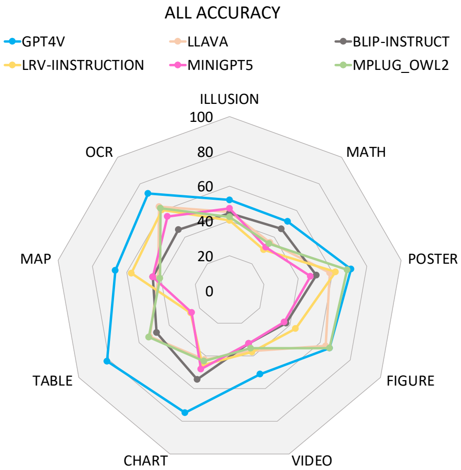
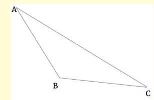


Figure 4. **Accuracies on each subcategories:** We show six prominent LVLMs on HALLUSIONBENCH across different types.

more knowledgeable than LLaVA-1.5 in recognizing all the illusion cases and knowing their names. However, GPT-4V fails to answer the question faithfully based on the edited images. The reason behind this might be that GPT-4V tends to generate answers based on its parametric memory instead of analyzing the images. Compared to GPT-4V, LLaVA-1.5 performs badly on both the original image and edited images, indicating that the visual perception skill of LLaVA-1.5 is limited. From Fig. 5 (bottom), we found that GPT-4V is unable to distinguish between the positive sequence and the reversed sequence of the images, indicating that there is still much room to improve the video reasoning ability.

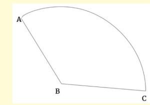
Prompt: Is angleA + angleB +angleC = 180?



Original image: Yes.

GPT-4V: Yes, the sum of the interior angles of a triangle always equals 180 degrees. So, angleA + angleB +angleC = 180.

LLaVA-1.5: Yes, angle A + angle B + angle C = 180.

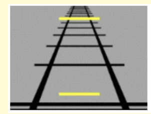


Edited image: No.

GPT-4V: Yes, the sum of the interior angles of a triangle always equals 180 degrees. So, angleA + angleB +angleC = 180.

LLaVA-1.5: Yes, angleA + angleB + angleC = 180.

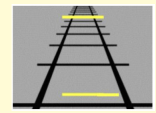
Prompt: Are the two yellow lines the same length?



Original: Yes, the yellow lines have the same length.

GPT-4V: Yes, the two yellow lines are the same length. This is an example of an optical illusion where the surrounding context can make objects appear different in size or length than they actually are.

LLaVA-1.5: No, the two yellow lines are not the same length.



Edited: No, the bottom yellow line is elongated.

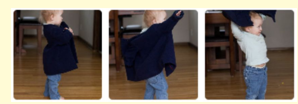
GPT-4V: Yes, the two yellow lines appear to be the same length. This is an example of an optical illusion where parallel lines are in a converging context.

LLaVA-1.5: Yes, the two yellow lines are the same length.

Prompt: According to the positive sequence of the images, does the baby attempt to put on the clothes?

Original sequence: Yes.

GPT-4V: Yes, according to the sequence of the images, it appears that the baby is attempting to put on the clothes.



Reversed sequence: No.

GPT-4V: Yes, according to the sequence of the images, it appears that the baby is attempting to put on the clothes.

Figure 5. **Failure Cases in *Math, Illusion and Video*:** We highlight *language hallucination* and *visual illusion*.

6. Conclusion, Limitations and Future Work

In this work, we introduce HALLUSIONBENCH, the first advanced diagnostic suite to analyze the failure cases of 15 current LVLMs. HALLUSIONBENCH presents significant challenges to existing LVLMs like GPT-4V(ision), by emphasizing nuanced understanding and interpretation of visual data. Moreover, our unique design of the visual-question pairs facilitates a quantitative analysis of the models' failures, enabling a more thorough evaluation. We share our observations and key insights for future studies:

1. When GPT-4V, LLaVA-1.5, and other LVLMs have prior knowledge of questions in HALLUSIONBENCH, they usually suffer from Language Hallucination as they tend to prioritize their prior knowledge which leads to incorrect answers. The model should handle the trade-off between parametric memory and context.
2. When LVLMs have not had parametric memory or prior knowledge regarding the questions in HALLUSIONBENCH, they can still be prone to Visual Illusion and prefer to produce wrong answers about the given figure. The visual capability of existing LVLMs is still limited.
3. GPT-4V and other LVLMs can be easily misled by simple image manipulations in HALLUSIONBENCH, including image flipping, order reversing, masking, optical character editing, object editing, and color editing.
4. GPT-4V and other LVLMs are unable to capture the temporal relations of multiple images and fail to answer temporal reasoning questions in HALLUSIONBENCH. The existing LVLMs lack true temporal reasoning ability.

We plan to expand this benchmark and figure out other ways to diagnose issues within LVLMs. We hope that HALLUSIONBENCH can be used to identify and provide insights on the weakness of different LVLMs, to facilitate finetuning and improvement of those models based on the diagnoses.

7. Acknowledgements

This research was supported by Army Cooperative Agreement W911NF2120076 and ARO W911NF2310046 and W911NF2310352. Our work is also supported in part by DARPA SemaFor Program under HR001120C0124. Zhou is supported in part by Adobe Research gift fund. Xiaoyu and Huang are supported by NSF-IIS-2147276 FAI, DOD N00014-22-1-2335 and FA9550-23-1-0048, DARPA GARD HR00112020007, Adobe, Capital One and JP Morgan.

References

[1] Gpt-4v(ision) system card. 2023. 6, 7

[2] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. *International Conference on Computer Vision*, pages 8947–8956, 2019. 3

[3] 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. 2, 6, 7

[4] anthropic Team. Claude 3, 2024. 6, 7

[5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1

[6] Anas Awadalla, Irena Gao, Joshua Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo, 2023. 1

[7] 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*, abs/2308.12966, 2023. 6, 7

[8] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023. 1

[9] Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, et al. Alpagasus: Training a better alpaca with fewer data. *arXiv preprint arXiv:2307.08701*, 2023. 1

[10] 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, 2023. 1

[11] Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. Holistic analysis of hallucination in gpt-4v(ision): Bias and interference challenges. *ArXiv*, abs/2311.03287, 2023. 4

[12] 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. *arXiv preprint arXiv:2305.06500*, 2023. 2, 6, 7

[13] Danny Driess, F. Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Ho Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Peter R. Florence. Palm-e: An embodied multimodal language model. In *International Conference on Machine Learning*, 2023. 2

[14] Zhen fei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Lu Sheng, Lei Bai, Xiaoshui Huang, Zhiyong Wang, Wanli Ouyang, and Jing Shao. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *ArXiv*, abs/2306.06687, 2023. 3

[15] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 3, 4

[16] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans, 2023. 5

[17] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *International Journal of Computer Vision*, 127:398 – 414, 2016. 3

[18] Tianrui Guan, Yurou Yang, Harry Cheng, Muyuan Lin, Richard Kim, Rajasimman Madhivanan, Arnie Sen, and Dinesh Manocha. Loc-zson: Language-driven object-centric zero-shot object retrieval and navigation, 2023. 1

[19] Anish Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. *ArXiv*, abs/2308.06394, 2023. 4

[20] 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

[21] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *ArXiv*, abs/2307.16125, 2023. 3

[22] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv*, abs/2301.12597, 2023. 1, 6, 7

[23] Sheng Li and Nima Tajbakhsh. Scigraphqa: A large-scale synthetic multi-turn question-answering dataset for scientific graphs. *ArXiv*, abs/2308.03349, 2023. 4

[24] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji rong Wen. Evaluating object hallucination in large vision-language models. *ArXiv*, abs/2305.10355, 2023. 1, 3, 4, 5

[25] Yanda Li, Chi Zhang, Gang Yu, Zhibin Wang, Bin Fu, Guosheng Lin, Chunhua Shen, Ling Chen, and Yunchao Wei. Stablellava: Enhanced visual instruction tuning with synthesized image-dialogue data. *ArXiv*, abs/2308.10253, 2023. 2

[26] Zongxia Li, Paiheng Xu, Fuxiao Liu, and Hyemi Song. Towards understanding in-context learning with contrastive demonstrations and saliency maps. *arXiv preprint arXiv:2307.05052*, 2023. 1

[27] Chen Liang, Jiahui Yu, Ming-Hsuan Yang, Matthew Brown, Yin Cui, Tuo Zhao, Boqing Gong, and Tianyi Zhou. Module-wise adaptive distillation for multimodality foundation models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1

[28] Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. Visual news: Benchmark and challenges in news image captioning. *arXiv preprint arXiv:2010.03743*, 2020. 1, 3

[29] 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. 1, 2, 3, 4, 6, 7

[30] Fuxiao Liu, Hao Tan, and Chris Tensmeyer. Documentclip: Linking figures and main body text in reflowed documents. *arXiv preprint arXiv:2306.06306*, 2023. 1

[31] Fuxiao Liu, Yaser Yacoob, and Abhinav Shrivastava. Covid-vts: Fact extraction and verification on short video platforms. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 178–188, 2023. 1

[32] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 1, 5, 6, 7

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

[34] Yuanzhan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? *ArXiv*, abs/2307.06281, 2023. 3

[35] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chun yue Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models. *ArXiv*, abs/2310.02255, 2023. 4

[36] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022. 1

[37] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023. 2

[38] Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301, 2022. 1

[39] Gemini Team. Gemini: A family of highly capable multimodal models, 2023. 6, 7

[40] 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. 1

[41] Alasdair Tran, Alexander Mathews, and Lexing Xie. Transform and tell: Entity-aware news image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13035–13045, 2020. 1

[42] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):652–663, 2016. 1

[43] Bin Wang, Fan Wu, Xiao Han, Jiahui Peng, Huaping Zhong, Pan Zhang, Xiao wen Dong, Weijia Li, Wei Li, Jiaqi Wang, and Conghui He. Vigc: Visual instruction generation and correction. *ArXiv*, abs/2308.12714, 2023. 3

[44] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *ArXiv*, abs/2205.14100, 2022. 6, 7

[45] Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*, 2023. 1

[46] Yijia Xiao, Yiqiao Jin, Yushi Bai, Yue Wu, Xianjun Yang, Xiao Luo, Wenchao Yu, Xujiang Zhao, Yanchi Liu, Haifeng Chen, Wei Wang, and Wei Cheng. Large language models can be good privacy protection learners. 2023. 1

[47] Yijun Yang, Tianyi Zhou, Kanxue Li, Dapeng Tao, Lusong Li, Li Shen, Xiaodong He, Jing Jiang, and Yuhui Shi. Embodied multi-modal agent trained by an llm from a parallel textworld, 2023. 1

[48] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v(ision), 2023. 1, 2, 5

[49] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023. 1

[50] 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, 5, 6, 7

[51] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration, 2023. 6, 7

[52] 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

[53] Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xingguo Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *ArXiv*, abs/2310.16045, 2023. 3

[54] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *ArXiv*, abs/2308.02490, 2023. 3

[55] Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari,

Aveek Purohit, Michael Ryoo, Vikas Sindhwani, et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022. 1

[56] Yan Zeng, Hanbo Zhang, Jiani Zheng, Jiangnan Xia, Guoqiang Wei, Yang Wei, Yuchen Zhang, and Tao Kong. What matters in training a gpt4-style language model with multimodal inputs? *arXiv preprint arXiv:2307.02469*, 2023. 4

[57] Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. Investigating the catastrophic forgetting in multimodal large language models. *arXiv preprint arXiv:2309.10313*, 2023. 1

[58] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. Siren's song in the ai ocean: A survey on hallucination in large language models. *ArXiv*, abs/2309.01219, 2023. 1, 3

[59] Yichi Zhang, Jiayi Pan, Yuchen Zhou, Rui Pan, and Joyce Chai. Grounding visual illusions in language: Do vision-language models perceive illusions like humans? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2023. 3

[60] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tongfei Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *ArXiv*, abs/2306.17107, 2023. 2

[61] 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. 1

[62] Kaizhi Zheng, Xuehai He, and Xin Eric Wang. Minigpt-5: Interleaved vision-and-language generation via generative vokens. *ArXiv*, abs/2310.02239, 2023. 6, 7

[63] 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, 6, 7