# VisDiaHalBench: A Visual Dialogue Benchmark For Diagnosing Hallucination in Large Vision-Language Models

**Qingxing Cao**[1] **Junhao Cheng**[1] **Xiaodan Liang**[1,2,3] **Liang Lin**[4]*
[1]Shenzhen Campus of Sun Yat-sen University [2]MBZUAI
[3]DarkMatter AI Research [4]Sun Yat-sen University
caoqx8@mail.sysu.edu.cn, linliang@ieee.org

## Abstract

Despite the significant success of large vision-language models (LVLMs), some studies have revealed that LVLMs suffer from the hallucination problem, where the LVLMs' response contains descriptions of non-existent objects. Although various benchmarks have been proposed to investigate this problem, they mostly focus on single-turn evaluation and overlook the hallucination raised by textual inputs. To investigate the hallucination problem of LVLMs when given long-term misleading textual history, we propose a novel visual dialogue hallucination evaluation benchmark VisDiaHalBench. The benchmark consists of samples with five-turn questions about an edited image and its original version. VisDiaHalBench differs from previous hallucination benchmarks in the following three points: 1) The questions and answers are unambiguously grounded by annotated scene graphs. 2) The images are uncommonly edited to inspect the visual model and common-object hallucination in LLMs. 3) The carefully designed dialogue refers a same object in different turns to assess the image consistency and influence of history for LVLMs. The detailed analysis of several state-of-the-art LVLMs across image consistency, visual understanding, history influence, and other dimensions reveals their substantial performance gap with single-turn VQA tasks. The benchmark is released in: https://github.com/qingxingcao/VisDiaHalBench

## 1 Introduction

Large language models (LLMs) (Chung et al., 2022; Touvron et al., 2023; Chiang et al., 2023; Achiam et al., 2023) have shown exceptional capabilities in semantic understanding, reasoning, and commonsense utilization. To expand their capabilities into the vision domain, recent research integrated vision-pretrained models with LLMs, resulting in Large Vision-Language Models (LVLMs) (Dai et al., 2023; Liu et al., 2023c; Zhu et al., 2023; Li et al., 2023a; Achiam et al., 2023). Through instruction finetuning on text-image paired data, LVLMs achieved great performance on a variety of multimodal tasks that require a vast range of skills. Despite the recent success, LVLMs also suffer from the hallucination problem, where the responses are non-factual or contradictory to the given inputs. To study this problem, recent works (Yifan Li and Wen, 2023; Wang et al., 2023; Liu et al., 2023b,a) have proposed different benchmarks and methods to diagnose hallucination within LVLMs. However, while previous works have shown that LLM will over-commit to early mistakes and leading to snowballing hallucination (Zhang et al., 2023; Yao et al., 2023), these studies in multi-modal domains mostly focus on the visual side, neglecting to fully investigate the textual inputs, such as misleading dialogue history or prior responses.

Given that LVLMs are built upon LLMs, there is a strong likelihood that LVLMs inherit the similar hallucination problem from LLMs. As shown in the second example of Figure 1, GPT-4V correctly answers the relation of the frisbee but subsequently hallucinates about a non-existent "white frisbee" given a follow-up similar question.

To analyze hallucinations in LVLMs when faced with multi-turn visual and textual inputs, we introduce VisDiaHalBench, a novel visual dialogue benchmark grounded by image scene graphs. Each sample in VisDiaHalBench contains an edited image and its original version to evaluate LVLMs' performance with unseen images; A scene graph to ground questions-answers for unambiguous evaluation; And a five-turn coherent dialogue with last turn references to investigate image consistency and history influence of LVLMs. For example, a correct answer about the edited brown frisbee will be an incorrect descrip-

* Corresponding author.



*Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12161–12176

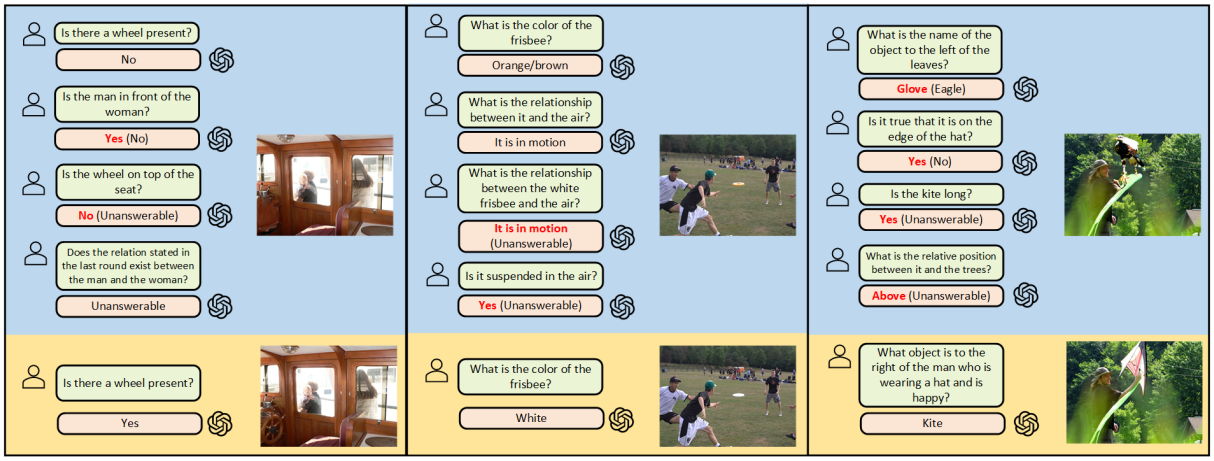


Figure 1: Three examples of our proposed VisDiaHalBench and GPT-4V results. Each sample contains five questions, where the first four questions query the edited image and the last one queries the original image. Incorrect answers are marked in red, while the groundtruth answers are provided in parentheses.

tion of the original white frisbee, which may affect LVLMs when answering the last question.

To construct the benchmark, we leverage the normalized scene graphs and corresponding images from the GQA (Hudson and Manning, 2019) dataset to construct VisDiaHalBench. For each image, we first edit the GQA image as well as the scene graph by removing or changing an object. Then in each turn, we sample a question type and scene graph paths based on the last turn object, attribute, or relation. With the grounded path and predetermined answers, we employ GPT-4 (Achiam et al., 2023) to generate a natural language question. We further benchmark several SOTA LVLMs and conduct comprehensive studies of their hallucination results. Our analysis reveals that current LVLMs still suffer from insufficient capabilities in vision understanding and handling complex dialogue history.

Our contribution can be summarized as follows: 1) we introduce the first visual dialogue benchmark focusing on diagnosing hallucinations raised by visual and long-term textual inputs. 2) We conduct comprehensive studies on SOTA LVLMs to investigate their image consistency, history handling, and other relevant aspects. We hope our benchmark and analysis can shed light on further research in addressing the hallucination problem in both language and vision domains.

## 2 Related Works

### 2.1 Large Visual-Language Models

With the success of the large language models (LLMs) (Chung et al., 2022; Touvron et al., 2023; Chiang et al., 2023; Achiam et al., 2023), some works (Alayrac et al., 2022; Li et al., 2023b) attempt to leverage the LLMs' powerful capability of semantic reasoning in vision domains, resulting Large Visual-Language Models(LVLMs). These works bridge the vision tokens to LLMs by only training a projector. Most recent works (Dai et al., 2023; Liu et al., 2023c; Zhu et al., 2023; Li et al., 2023a; Liu et al., 2023b) employ instruction tuning, where the model can be trained on different multi-modal tasks in a unified manner, and achieve impressive results on various downstream tasks.

### 2.2 LVLMs Hallucination and Evaluation

Despite the success of LVLMs, recent works suggest they suffer from the hallucination problem (Liu et al., 2023a; Zhou et al., 2023), where the response has inaccurate descriptions of a given image. Recent works (Yifan Li and Wen, 2023; Wang et al., 2023; Liu et al., 2023b) have proposed different methods and benchmarks to deeply inspect this phenomenon. POPE (Yifan Li and Wen, 2023), HaELM (Wang et al., 2023) and GAVIE (Liu et al., 2023b) differently prompt a LLM to score LVLMs' response. Liu et al. (2023a) and Cui et al. (2023) manually edit visual inputs and label question-answering pairs to assess LVLMs. Different from these works, we propose a visual dialogue benchmark to thoroughly investigate the hallucination in LVLMs from both vision and long-term text aspects.

### 2.3 Visual Dialogue

Given an image, a dialogue history, and a question about the image, Visual dialogue (VD) (Das et al., 2017) requires an agent to generate an ac-

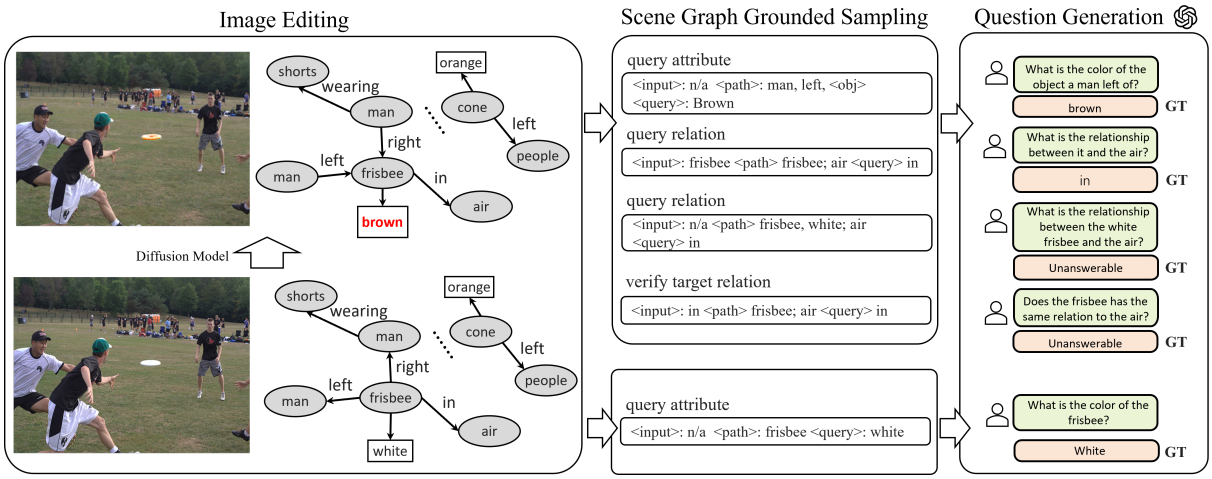


Figure 2: The construction pipeline for our proposed VisDiaHalBench. We first edit the image and corresponding scene graph. Then sample questions type and scene graph path based on previous questions. Lastly, we prompt GPT-4 to generate natural language questions.

curate response based on the image and informed by the dialogue history. Recent advance in pre-trained visual-language models (Su et al., 2020; Li et al., 2022a) also inspired related models in VD, such as VisDial-BERT (Murahari et al., 2020), VD-BERT (Wang et al., 2020) ICMU (Chen et al., 2022b), UTC (Chen et al., 2022a), BLIP (Li et al., 2022a) and AlignVD (Chen et al., 2022c). Unlike the single-turn Visual Question Answering, the challenge of VD lies in the need to address referential and ambiguity issues present in the historical dialogue (Chen et al., 2020). Previous studies explored how to increase the dialogue models' robustness. Yu and Rieser (2023) study the adversarial robustness of visual dialogue models by attacking the vulnerable words within questions or history. Kang et al. (2023) propose a Generative Self-Training (GST) method that can automatically generate dialogue samples for data augmentation and model training. Our work also aims to evaluate model robustness concerning dialogue history but employ the history generated by LVLMs themselves without modification, studying a more general phenomenon that severely hinders the further application of LVLMs.

## 3 VisDiaHalBench

We introduce VisDiaHalBench, a new visual dialogue benchmark to analyze hallucination in LVLMs under different visual and textual inputs. Based on the normalized scene graph and image in the GQA dataset, each sample contains an edited image, a five-turn dialogue, and grounding scene graph paths. The questions are associated with the edited object in different aspects, including queries about the uncommon edited results, pronoun-based reference, and misleading history.

To construct the benchmark, we first edit the image by randomly changing an object's attribute, swapping it with another object, or removing it from the image. Then at each turn, we sample a question type and scene graph path based on the edited object or previous questions. Lastly, we prompt GPT-4 (Achiam et al., 2023) to generate a natural language question strictly grounded to the supporting paths and ground truth answer. The overall generation pipeline is shown in Figure 2.

To ensure the unambiguity of generated questions and answers, we rely on sampling scene graph path that can uniquely identify an object. Also, by designating the start or end object of a path, we can generate coherent multi-turn questions concerning the last-turn context. To better illustrate the construction process, we first present our path sampling method and question type details, then describe each construction step.

### 3.1 Scene Graph Path Sampling

We generate unambiguous questions and answers grounded by paths in the scene graph. A scene graph path is a sequence of connected object nodes and relations. For example, a 0-hop path is just an object like "frisbee", a 1-hop path could be "(frisbee, left, man)", and a 2-hop path might extend to "(frisbee, left, man), (man, wearing, shorts)".

To sample a $n$-hop path, we begin at an object node, randomly select one of its relations, and traverse to the corresponding node $n$ times. Then we verify whether the path is unique in the

| Question Type | Path Example | Question Example | Input Reference | Output Reference |
|---|---|---|---|---|
| exist:<br>Query the existence of a object | [(frisbee,left,<obj>),<br>(<obj>,<name>,man)] | Is there a man that a frisbee on its left? | Object | Object (if answer is "Yes") |
| query object:<br>Query a object name | [(man,wearing,<obj>),<br>(<obj>,<name>,shorts)] | What is the man wearing? | Object | Object |
| query attribute:<br>Query a object's attribute | <input>: frisbee,<br>[(frisbee,<attribute>,yellow)] | What color is it? | Object | Object, Attribute |
| query relation:<br>Query relation of two objects | <input>: frisbee, [frisbee],<br>[air] | What is the relationship between it and the air? | Object | Object, Relation |
| verify attribute:<br>Query if object has a attribute | [(frisbee,<attribute>,yellow)]<br>Query: white | Is the frisbee white? | Object | Object, Attribute |
| verify relation:<br>Query if two objects have a relation | [(frisbee,<attribute>,white)],<br>[man] Query: left | Does the white frisbee to the left of the man | Object | Object, Relation |
| verify target attribute:<br>Query if object has a mentioned attribute | <input>: white, [frisbee] | Does the frisbee has the same color? | Attribute | Object |
| verify target relation:<br>Query if two objects have a mentioned relation | <input>: left, [cone], [people] | Does the cone and the people have the same relation? | Relation | Object |

Table 1: The details of all question types in VisDiaHalBench. Path and question examples give a possible scene graph path and corresponding question. Input and output reference represents the object or answer that can be used in this turn or passed to the next turn.

scene graph. We obtain all objects with the same name and check whether these objects have the same relation recursively. For example, we verify whether "(frisbee, left, <obj1>), (<obj1>, wearing, <obj2>)" is unique in the scene graph, such that the shorts can be unambiguously referred to as "the thing wearing by one a frisbee left of".

In our benchmark, we sample 1-hop and 2-hop when questions target object names without giving away the object name. For other questions, we sample 0-hop and 1-hop paths.

**Coherent path sampling** To construct coherent dialogue, we sample paths and question types that can accept the object or answer mentioned in the last turn, as shown in Table 1. If the last turn refers to an object, we sample a path starting from it and use pronouns to refer to this object. If the last turn has an attribute or relation, we first sample the answer "Yes/No", then sample a path that ends with an object that has or does not have this attribute/relation according to the answer.

**Sampling Retry** During the random sampling process, if a sample path is not unique, or the sampled query object has no attribute or relation to the query, we re-sample the whole dialogue until reach a predefined attempt limit. If the limit is reached, we discard this sample and continue to sample the next one.

### 3.2 Question Types

The questions in VisDiaHalBenchhave following eight types: "exist", "query object", "query attribute", "query relation", "verify attribute", "verify relation", "verify target relation" , "verify target attribute". These question types represent queries about the existence of an object, its name, attribute, or relations to another object. Or verify whether they possess certain attribute or relation. In each turn, we first sample a question type from its subset according to the turn number and last turn question. The details are shown in Table 1.

For the question that verifies whether the object has an attribute or relation, we need to sample an unrelated query target if the answer is "no". We select the name/attribute/relation that is similar to a correct one. Specifically, we first extract all object names, attributes, and relations in the dataset and form three dictionaries, then embed these words or phrases with GPT2-large (Lagler et al., 2013). Given a correct name/attribute/relation, we select its similar words by computing its embedding distance with other names/attributes/relations respectively. We randomly select one from the top 10

to obtain a similar name/attribute/relation and exclude the top 5 to avoid retrieving synonym words. The retrieved words are used for questioning.

To balance the number of each question type, we set the probability of selecting different types according to their generated number. The probability for type $T$ is $1 - \frac{\text{number of questions type T}}{\text{number of total question}}$. Thus types with more samples will have a lower probability of being selected.

### 3.3 Image Editing

Since most of the current LVLMs achieve great performance on the GQA dataset, we edit the GQA image to explore whether the hallucination comes from the biases in training data. To obtain a visually reasonable image, we only modify a certain object in one of three different ways: remove it from the image, swap it with another object, and change its color. We randomly choose the edition type and target so that the image can be counterfactual for usual scenarios. First, we use the segmentation model SAM (Kirillov et al., 2023) to obtain objects' mask according to their scene graph bounding boxes. Subsequently, we employ various strategies for image editing.

**Remove object** To remove an object, we first select the object that can be uniquely determined by 0-2 hop path in the original scene graph, so that the following questions and answers are not ambiguous. For example, the answer for "Is there an umbrella in the image" remains the same "Yes" if there exists multiple umbrellas but only one of them is removed.

Given the identifiable object, we retrieve its bounding boxes from the annotated scene graph and instruct the pretrained image editing model IP-adapter (Ye et al., 2023) to remove it from the image. Specifically, we utilize the inpainting mode in the IP-Adapter to remove the corresponding objects. By setting the prompt to "empty" and the negative prompt to "$n_{obj}$" which represents the name of the object to be removed, we can obtain images without the specified objects. Lastly, we modify the scene graph correspondingly by removing this object and all relations to other objects.

**Change object** We change the whole selected object by replacing it with another object in the GQA dataset such that the object is recognizable but unseen in the image. Specifically, we extract all the object names in the dataset and randomly select a similar one based on GPT-2 embedding, the same as the method described in Section 3.2. With the candidate object name, we randomly select an object that has this name from all the images in GQA dataset as the reference object.

We change the original object to the reference object through Anydoor (Chen et al., 2023) method. We input the two masks and corresponding images to the Anydoor and copy the reference object to the original object mask. Lastly, we correspondingly modify the scene graph by setting the object name and the attributes to the reference image but keeping the relations unchanged.

**Modify attribute** Since the GQA annotation has various attribute types, including actions that are only applicable to some objects, adding an attribute does not necessarily replace an existing one. Thus, to obtain a reasonable attribute and avoid ambiguity, we only change the object color. Similar to the object name, we first obtain all attributes in the GQA dataset, then filter the color using the matplotlib (Hunter and Dale, 2007) colors package and form a color set. Then we select the object that has a color attribute and replace its color attribute with another different one randomly selected from the color set.

Given the original color, reference color, and the object mask obtained by SAM, we instruct the Blended latent diffusion (Avrahami et al., 2023) model with prompt "$c_{new} + n_{obj}$" to perform the editing. Where $c_{new}$ represents the new color selected and $n_{obj}$ represents the object's name. Finally, we change the object in the scene graph by removing its original color and adding the reference color to its attribute set.

### 3.4 Dialogue Sampling

After image editing, we generate dialogue questions to investigate different aspects of LVLM hallucination. Each dialogue contains five turn questions that query the visual information related to the edited object. At each turn, we sample a question type then sample scene graph paths and answer. We constrain sampling strategy in different ways to inspect different aspects of LVLMs. The details of each turn is shown in the following.

**Turn 1** : The first question directly queries the edited part of the object to inspect the vision model of LVLMs. For "remove object", we query whether the object "exist" by sampling a 0 or 1-hop path ended with the edited object from the original scene graph. For "change object", we use

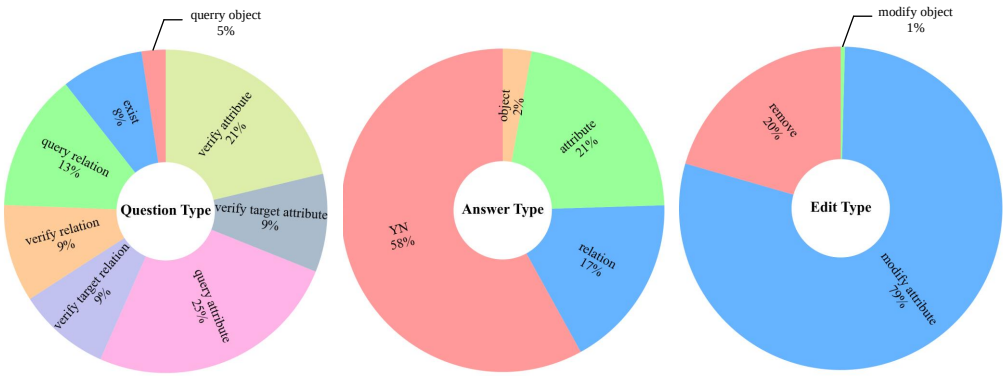


Figure 3: The data distribution of VisDiaHalBench, sorted by (a)question type (b)answer type (c)edit type

| | exist | query object | query attribute | query relation | verify attribute | verify relation | verify target attribute | verify target relation | Total |
|---|---|---|---|---|---|---|---|---|---|
| remove | 2054 | 168 | 485 | 612 | 240 | 623 | 268 | 685 | 5135 |
| modify attribute | 0 | 390 | 5889 | 2809 | 5056 | 1800 | 2186 | 1630 | 19760 |
| modify object | 0 | 44 | 9 | 17 | 19 | 15 | 1 | 0 | 105 |
| Total | 2054 | 602 | 6383 | 3438 | 5315 | 2438 | 2455 | 2315 | 25000 |

Table 2: Number of samples for different question types and edit types in VisDiaHalBench.

"query object" to query the name of the changed object, and sample 1 or 2-hop path from the edited scene graph to avoid directly refer the object with its name. For "modify attribute" edition, we randomly choose "query attribute" and "verify attribute" and sample 0 or 1-hop scene graph path. If the question type is "verify attribute" and the randomly selected answer is "no", we used the original attribute to generate the question. For example, if we change a frisbee from white to brown, we will have questions: "Is the frisbee white?".

**Turn 2** : This question queries the other part of the image related to the turn 1 question to study the dialogue ability of LVLMs. We refer to the last turn object or answer with a pronoun and sample scene graph path with method in Section 3.1

**Turn 3** : We mention the edited object with its original name or attribute to generate misleading questions that can not be answered. This turn is designed to explore whether the LVLMs hallucinate the answer or consistent with turn 1. For example, given a frisbee changed from white to brown, "What is the white frisbee related to the air?" has no answer since no white frisbee in the image. LVLMs should not answer if it is consistent with Turn 1 where it knows the frisbee is brown.

**Turn 4** : Similar to turn 2, the turn 4 sample path is based on the last turn question. However, since the last turn object and answer are not valid, this question is also unanswerable. We employ this question to further study the robustness of LVLMs in handling corrupted history.

**Turn 5** : We employ the question identical to turn 1 but provide the unedited image. Current LVLMs should be able to easily answer this GQA-like question. We study the performance gap between answering it directly or in dialogue to evaluate how much the dialogue history affects LVLMs.

### 3.5 Question Generation

Given the object path, related attributes, and relations in each turn, we can obtain the unambiguous answer grounding by supporting paths. We obtain the natural language question by prompting GPT-4 to generate it that strictly grounding to the supporting triplets and answer. The specific prompt can be found in the appendix B. We resample the question if it contains words of last turn reference, or not contains the grounded attributes and name.

### 3.6 Dataset Statistic

Following the aforementioned editing method, we edited 5000 images from GQA and subsequently generated 5000 diverse visual dialogues, each dialogue consists of 5 turns of conversation, totaling 25,000 visual question-answers. As shown in figure 3, our dataset includes 8 types of question types, 4 types of answer types, and 3 types of edit types. Table 2 enumerates the specific quantities for different question types and edit types.

| Model | Turn 1 | | Turn 2 | | Turn 3 | | Turn 4 | | Turn 5 | | Average | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | EM | F1 | EM | F1 | EM | F1 | EM | F1 | EM | F1 | EM | F1 |
| *w/ image:* | | | | | | | | | | | | |
| BLIP2 | 19.24 | 25.21 | 0.04 | 17.42 | 0.0 | 35.47 | 0.0 | 32.51 | 0.0 | 18.26 | 3.85 | 25.77 |
| Cheetah | 19.96 | 27.55 | 19.56 | 30.14 | 0.25 | 25.01 | 0.22 | 23.27 | 18.98 | 25.89 | 11.80 | 35.73 |
| MiniGPT4 | 5.31 | 13.22 | 26.36 | 36.89 | 0.07 | 28.72 | 0.36 | 27.59 | 71.74 | 79.02 | 20.77 | 37.08 |
| LRV-instruct | 17.78 | 26.86 | 24.36 | 40.18 | 0.07 | 33.63 | 0 | 35.48 | 22.50 | 34.23 | 12.94 | 33.66 |
| HalluDoctor | 39.95 | 49.22 | 34.18 | 44.57 | 0.0 | 27.02 | 0.23 | 25.86 | 35.79 | 43.21 | 22.03 | 37.98 |
| LLaVA | 26.32 | 32.52 | 38.72 | 54.58 | 0.07 | 25.91 | 0.25 | 17.97 | 54.29 | 57.67 | 23.93 | 37.99 |
| GPT-4V | 25.51 | 38.23 | 33.50 | 47.81 | 38.72 | 55.61 | 17.83 | 35.93 | 46.11 | 55.45 | 33.01 | 48.26 |
| Q-Probing | 21.49 | 31.83 | 27.40 | 37.66 | 3.00 | 31.10 | 1.65 | 28.25 | 61.57 | 70.06 | 23.02 | 39.78 |
| *w/ scene graph:* | | | | | | | | | | | | |
| LLaVA | 35.72 | 41.22 | 53.74 | 61.78 | 10.06 | 35.36 | 1.57 | 23.02 | 88.26 | 90.42 | 37.87 | 50.36 |
| GPT-4V | 82.22 | 85.34 | 63.97 | 70.75 | 32.25 | 65.34 | 20.09 | 43.58 | 99.07 | 99.39 | 59.12 | 72.88 |
| Human | 89.21 | 93.75 | 88.24 | 94.55 | 100 | 100 | 100 | 100 | 99 | 99 | 95.29 | 97.46 |

Table 3: Evaluation results of different LVLMs on VisDiaHalBench.

# 4 Experiments

## 4.1 Methods

**Q-Probing** Previous research has shown that LLM may hallucinate due to early mistakes (Zhang et al., 2023). To alleviate this problem, we propose a simple baseline approach "Q-Probing" that instructs LVLMs to probing the relevant question to reduce the hallucination. Specifically, the Q-Probing approach involves instructing the LVLM to generate several closely related questions based on the image and current question. These generated questions act as reference points. During the question-answering phase, we prompt the LVLM with the instruction "Please consider these questions before answering", encouraging it to pay attention to additional crucial information and logical fallacies when formulating responses, thereby mitigating the phenomenon of hallucination. In the experimental section, we built our model based on MiniGPT-4 (Zhu et al., 2023) and demonstrated its effectiveness. Further details about the prompts can be found in appendix B.

**SOTA methods** We conduct an evaluation of several state-of-the-art LVLMs: BLIP2 (Li et al., 2023b), LRV-instruction (Liu et al., 2023b), HalluDoctor (Liu et al., 2024), MiniGPT4 (Zhu et al., 2023), Cheetah (Li et al., 2023a), LLaVA (Liu et al., 2023c), and GPT-4V on our benchmark. The implementation details are in appendix C.

## 4.2 Evaluation Metrics

We conduct a comprehensive evaluation of the baseline models, focusing on their performance in evidence retrieval and answer extraction. Following previous multimodal conversational dataset (Li et al., 2022b), we reported macro-average F1 in the word level and Exact Match (EM) to estimate the performance of answer extraction after extracting the model's output results through keyword filtering.

## 4.3 Main Results

Table 3 presents the performance of each model on VisDiaHalBench. All models exhibit subpar performance on average, the best-performing GPT-4 achieves 33.01 EM and 48.26 F1, far worse than its performance on GQA. This discrepancy indicates that our benchmark poses a new challenge for current LVLMs. All open-sourced models except Cheetah perform better in Turn 2 and Turn 5 compared to Turn 1. Since turn 2 and 5 query the unedited object, this result indicates that these LVLMs struggle with correctly perceiving the edited image. Furthermore, all models exhibit their worst performance in Turn 4, indicating that referring to a non-existent object, which misleads models in both the visual and textual domains, is the most challenging task for current LVLMs.

HalluDoctor and LRV-instruct are both based on MiniGPT4 and they greatly improve over baseline MiniGPT4 in perceiving edited objects, showing their effectiveness in handling object hallucination. However, their performance drop significantly in Turn 3-5. Especially in Turn 5, they perform much worse than the baseline MiniGPT4, showing the two single-turn hallucination mitigation may hinder LVLMs ability to handle multi-turn dialogues effectively.

The proposed "Q-Probing" outperforms both MiniGPT4 and LRV-instruct in average perfor-

| | Turn 1 | Turn 2 | Turn 3 | Turn 4 | Turn 5 | Average |
|---|---|---|---|---|---|---|
| LLaVA(Order 12345) | 26.32 | 38.72 | 0.07 | 0.25 | 54.29 | 23.93 |
| LLaVA(Order 34125) | 0 | 0.15 | 36.69 | 30.55 | 49.16 | 23.31 |
| GPT-4V(Order 12345) | 25.51 | 33.50 | 38.72 | 17.83 | 46.11 | 33.01 |
| GPT-4V(Order 34125) | 44.16 | 27.44 | 27.29 | 21.14 | 47.79 | 33.56 |

Table 4: The exact match of LVLMs in different dialogue orders. "Order 34125" represent the performance when first answer the unanswerable turn 3 and 4 questions without misleading turn 1 and turn 2 answers.

| Model | Turn 1 | w/ misleading history | | | | w/ history | | | | w/o history | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Non-existence object | | Original object | | Non-existence object | | Original object | | Original object | |
| | EM | EM | F1 | EM | F1 | EM | F1 | EM | F1 | EM | F1 |
| Cheetah | 19.96 | 0.18 | 20.22 | 18.58 | 23.42 | 0.27 | 27.31 | 19.08 | 27.49 | 53.67 | 65.44 |
| LLaVA | 26.32 | 0.14 | 30.32 | 32.60 | 45.17 | 0.05 | 13.44 | 62.04 | 71.29 | 73.02 | 82.72 |
| GPT-4V | 25.51 | 39.76 | 59.23 | 38.35 | 50.29 | 38.35 | 49.23 | 53.04 | 68.72 | 77.16 | 88.01 |

Table 5: The performance of LVLMs under different conditions. "w/ misleading history" means LVLMs correctly answer the Turn 1 question and "w/ history" means otherwise. Turn 3 and Turn 5 questions query a "Non-existence object" and the unedited "original object". "w/o history" shows the performance of LVLMs answering the Turn 5 question outside the dialogue.

mance. Compared with LRV-instruct, Q-Probing achieves large improvement in turn 1 and turn 5, whose questions query about the edited object and original object with misleading history. The results show that relevant questions can allow LVLMs to be more aware of additional logical and visual information and mitigate hallucinations.

### 4.4 More Analytical Study

**Hallucination with misleading questions** Turn 3 and 4 are designed to present LVLMs with misleading prompts that pose ambiguous questions. The objective was to assess whether LVLMs would recognize the flawed nature of these queries and respond appropriately with 'Unanswerable'. According to the results in table 3, there is a significant performance drop for all models in turns 3 and 4, suggesting that misleading prompts can easily lead to the generation of hallucinations, emphasizing the challenges associated with handling ambiguous queries accurately.

**Hallucination without misleading history** We evaluate the performance of LLaVA and GPT-4V with different orders in dialogue. Specifically, Since turn 2 and turn 4 questions refer to previous objects or answers, changing their order will lead to incorrect GT answers. Thus we change the order from the original 1 2 3 4 5 to 3 4 1 2 5, where the LVLMs firstly answer the original turn 3 and turn 4 questions. In this way, the LVLMs will not be affected by misleading history when answering the hardest unanswerable questions. The exact match (EM) results are shown in Table 4. It is shown that the GPT-4V can identify more answerable questions without misleading history, improving EM from 38.72 to 44.16. The unanswerable response also helps GPT-4V perceive edited objects, improving EM from 25.51 to 27.29.

**Visual understanding** To assess the effectiveness of the visual encoder, we conduct additional evaluations by providing LVLMs with scene graphs instead of images. The results, presented in the lower part of Table 3, reveal that LVLMs achieve superior performance compared to the image-based outputs. This indicates the bottleneck of current LVLMs may lie in the visual encoder.

**Image consistency** To inspect whether the LLM hallucinates regardless of the visual information, we evaluate whether a LVLM can generate consistent output about a same object. As Turn 1, 3, and 5 query the same object based on edited and original image, we evaluate the performance of LVLMs in Turn 3 and Turn 5, given that Turn 1 was answered correctly. As shown in Table 5 "w/ misleading history", if the LVLM correctly answers Turn 1, LVLM will has misleading history and tends to obtain better performance in turn 3(Non-existence object) but shows a significant drop on turn 5(original object), which contains a similar question but a different image. These findings suggest that LVLMs are influenced by their previous responses and tend to hallucinate answers without effectively utilizing visual inputs.

**Influence of history** To investigate whether the previous dialogue would affect the efficacy of the

| | Accuracy | Agreement Ratio |
|---|---|---|
| GT answer | 98% | 95.58% |

Table 6: Accuracy of groundtruth answer evaluated by human.

| | Accuracy | Agreement Ratio |
|---|---|---|
| Edition Correctness | 1.8 | 86% |
| Image Quality | 1.7 | 82% |

Table 7: Human evaluation of the edited images.

| | F1 | | | EM | | | Human | |
|---|---|---|---|---|---|---|---|---|
| | pearson | spearmanr | kendalltau | pearson | spearmanr | kendalltau | Accuracy | Agreement |
| LLaVA | 0.6723 | 0.5532 | 0.5031 | 0.7896 | 0.7896 | 0.7896 | 26% | 92.4% |
| GPT4-V | 0.6611 | 0.5123 | 0.4912 | 0.7527 | 0.7526 | 0.7527 | 44.8% | 88% |

Table 8: The human evaluation results of LVLMs and the correlation coefficient to the EM and F1.

current response of LVLMs, we evaluate the GQA-like question in turn 5 given the premise that turn 1 was answered correctly or incorrectly. As described above, if LVLMs answer correctly in Turn 1, the outcomes in Turn 5 with similar image will be affected. Additionally, we evaluate the performance of LVLMs when directly asking Turn 5 questions outside of the dialogue context. The impressive results, presented in Table 5, indicate that LVLMs perform well when directly asked Turn 5 questions without the influence of prior dialogue. The performance gap observed in these evaluations indicates the existence of hallucination induced by the dialogue history.

More study of LVLMs' performance on different edit types is in appendix A.

### 4.5 Human studies on VisDiaHalBench

**Generated images** To validate our generated images, we randomly select 50 images and ask two annotators to rate an edited image from 0 2, based on whether the edition follows the instructions and whether the image seems natural. The results are shown in Table 7. The agreement ratio represents the extent to which the rating provided by the two annotators matches. As shown in the table, the diffusion models can correctly edit the objects in most cases given the annotators have an average rating of 1.8 and 86% of rating agreement.

**Correctness of GT answers** We ask two human annotators to answer the questions for 50 sampled dialogues with a total of 250 questions. We present the annotator with possible color candidates to avoid answering with color synonyms. The results are shown in the last line of Table 3. Human-labeled answers achieve a 95.29% accuracy compared to the groundtruth answers. When querying object names, some human answers are synonyms of the groundtruth, leading to lower exact match (EM) but higher F1 scores. After labeling some samples, the human annotator became aware that Turn 3 and 4 consistently yield the fixed answer "unanswerable," resulting in a 100% EM. In conclusion, human annotator achieves F1 scores greater than 90 in all turns, indicating that the samples generated with the scene graph can be utilized to evaluate LVLMs.

We also evaluate whether the groundtruth answer is the same as human evaluation. The results are shown in Table 6. The "Accuracy" represents the correctness of the groundtruth answer evaluated by humans. The agreement ratio represents that two annotators have the same label for 95.58% of answers. The results align with the previous table, where groundtruth and human annotators reach an agreement in approximately 98% of the answers. These two tables show the correctness of generating questions and answers.

**Evaluation metrics** To study the correctness of the EM/F1 evaluation metrics, we conduct a human evaluation of the LLaVA and GPT4-V outputs. Specifically, we ask two annotators to score each answer in 50 dialogues with 0/1. The correlation coefficient between human and automatic evaluation metrics are shown in Table 8. The agreement ratio represents the extent to which the answers provided by the two annotators match. The table demonstrates a strong correlation between the automatic evaluation metrics and human evaluation.

## 5 Conclusion

In this work, we propose VisDiaHalBench, a novel visual dialogue benchmark specifically designed to investigate the hallucination phenomenon in Large Vision-Language Models (LVLMs). Our comprehensive analysis shows the weaknesses of current LVLMs in terms of image consistency and dialogue history, highlight the challenge of hallucination in vision-language understanding.

## 6 Limitations

Although we carefully designed the construction process and prompt to generate the natural language questions, some question texts are not fluent. Besides, although the question-answer is grounded by an annotated scene graph, an open-ended question can still have multiple answers with similar meanings. To better evaluate the answering accuracy objectively, the benchmark could have Yes/No questions only. The potential issue might include the misuse of the edited images might be misused for unexpected purposes.

## Acknowledgements

This work was supported in part by National Science and Technology Major Project (No.2021ZD0111601), National Natural Science Foundation of China (No.62325605), National Key R&D Program of China under Grant No. 2020AAA0109700, China Postdoctoral Science Foundation under Grant Number 2023M744001, Guangdong Outstanding Youth Fund (Grant No. 2021B1515020061) Mobility Grant Award under Grant No. M-0461, Shenzhen Science and Technology Program (Grant No. GJHZ20220913142600001), Nansha Key RD Program under Grant No.2022ZD014, CAAI-Huawei MindSpore Open Fund. We thank MindSpore for the partial support of this work, which is a new deep learning computing framwork[1].

## References

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

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *NeurIPS*.

Omri Avrahami, Ohad Fried, and Dani Lischinski. 2023. Blended latent diffusion. *ACM Transactions on Graphics (TOG)*, 42(4):1–11.

[1]https://www.mindspore.cn/

Cheng Chen, Zhenshan Tan, Qingrong Cheng, Xin Jiang, Qun Liu, Yudong Zhu, and Xiaodong Gu. 2022a. Utc: a unified transformer with inter-task contrastive learning for visual dialog. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 18103–18112.

Feilong Chen, Xiuyi Chen, Shuang Xu, and Bo Xu. 2022b. Improving cross-modal understanding in visual dialog via contrastive learning. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7937–7941. IEEE.

Feilong Chen, Duzhen Zhang, Xiuyi Chen, Jing Shi, Shuang Xu, and Bo Xu. 2022c. Unsupervised and pseudo-supervised vision-language alignment in visual dialog. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4142–4153.

Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. 2023. Anydoor: Zero-shot object-level image customization. *arXiv preprint arXiv:2307.09481*.

Xiaofan Chen, Songyang Lao, and Ting Duan. 2020. Multimodal fusion of visual dialog: A survey. In *Proceedings of the 2020 2nd International Conference on Robotics, Intelligent Control and Artificial Intelligence*, pages 302–308.

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

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

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

W Dai, J Li, D Li, AMH Tiong, J Zhao, W Wang, B Li, P Fung, and S Hoi. 2023. Instructblip: towards general-purpose vision-language models with instruction tuning. arxiv. *Preprint posted online on June*, 15:2023.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh,

and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *CVPR*.

John Hunter and Darren Dale. 2007. The matplotlib users guide. *Matplotlib 0.90. 0 users guide*.

Gi-Cheon Kang, Sungdong Kim, Jin-Hwa Kim, Donghyun Kwak, and Byoung-Tak Zhang. 2023. The dialog must go on: Improving visual dialog via generative self-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6746–6756.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Klemens Lagler, Michael Schindelegger, Johannes Böhm, Hana Krásná, and Tobias Nilsson. 2013. Gpt2: Empirical slant delay model for radio space geodetic techniques. *Geophysical research letters*, 40(6):1069–1073.

Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Hanwang Zhang, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, and Yueting Zhuang. 2023a. Fine-tuning multimodal llms to follow zero-shot demonstrative instructions. *arXiv preprint arXiv:2308.04152*, 3.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models.

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

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022a. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*.

Yongqi Li, Wenjie Li, and Liqiang Nie. 2022b. Mmcoqa: Conversational question answering over text, tables, and images. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4220–4231.

Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2023a. Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multi-modality models. *arXiv preprint arXiv:2310.14566*.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023b. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023c. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. 2020. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. In *European Conference on Computer Vision*, pages 336–352. Springer.

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

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations. In *ICLR*.

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

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, Jitao Sang, and Haoyu Tang. 2023. Evaluation and analysis of hallucination in large vision-language models.

Yue Wang, Shafiq Joty, Michael R Lyu, Irwin King, Caiming Xiong, and Steven CH Hoi. 2020. Vd-bert: A unified vision and dialog transformer with bert. *arXiv preprint arXiv:2004.13278*.

Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, and Li Yuan. 2023. Llm lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469*.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Kun Zhou Jinpeng Wang Wayne Xin Zhao Yifan Li, Yifan Du and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Lu Yu and Verena Rieser. 2023. Adversarial textual robustness on visual dialog. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3422–3438, Toronto, Canada. Association for Computational Linguistics.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2023. How language model hallucinations can snowball.

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

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

## A Performance analysis on edit types

Table 15 shows the LVLMs' performance on different types of edited image. It is shown that LVLMs perform worse on "modified objects" images because the questions and edited images are harder than the other two edit types.

Removing an object will impaint the region with its surrounding background, which makes the image still plausible in a natural context. Furthermore, querying its existence is the only valid question for a removed object. Thus, an LVLM only needs to answer yes/no for this edit type and has a higher answering accuracy.

Modifying attributes will change an object's color. Although the color may be unusual for the object, the object itself is reasonable for the scene. Thus, an LVLM can correctly answer some of these questions.

Modifying an object will replace the original object with another one that does not fit the scene. Additionally, the answering vocabulary for object names is much larger than for colors. The difficulty in both the image and the question leads to lower performance. The results also suggest that LVLMs may pay more attention to the commonsense inside the LVLM instead of visual tokens.

## B Instruct prompts

Our prompt for question generation is listed in table 9. The prompt for Q-Probing to generate relevant questions and answer original questions are listed in Table 10 and Table 11 respectively.

## C Model details in our benchmark

**Cheetah** integrates visual information and language understanding to handle zero-shot tasks, meaning it can execute tasks without prior examples. To enhance the model's performance in understanding complex visual and verbal instructions, Cheetah incorporates the VPG-C module, which is specifically designed to capture and supplement detail information. Moreover, Cheetah fine-tunes the VPG-C through a synthetic discriminative training strategy, thereby reducing the reliance on labeled demonstration data. The model variant evaluated in our experiments is "cheetah-llama-2-7b". The evaluation process is completed over a period of 50 GPU hours, utilizing one NVIDIA GeForce RTX 3090 GPU with 25GB of memory.

**LLaVA** propose a framework that fuses the capabilities of a visual encoder, specifically the ViT-L/14 from CLIP(Radford et al., 2021), with the language decoder abilities of LLaMA(Touvron et al., 2023). This integration is achieved using an intermediary fully-connected (FC) layer. The training process begins with the FC layer being trained in isolation on a dataset of 595,000 image-text pairs, with the pre-existing parameters of both the visual encoder and the language model remaining unchanged. Subsequent to this phase, a fine-tuning step is conducted where both the FC layer and the language model are jointly optimized. This fine-tuning employs a specialized dataset consisting of 158,000 pairs of instructional vision-language data. The model variant evaluated in our experiments is "LLaVA-v1.5-13b.". The evaluation process is completed over a period of 30 GPU hours, utilizing one NVIDIA GeForce RTX 3090 GPU with 25GB of memory.

**MiniGPT-4** employs a fully-connected (FC) layer to facilitate communication between a visual encoder and a text encoder. The initial training phase involves the FC layer learning from a dataset with 5 million image-text pairs. Following this, the model undergoes fine-tuning with a targeted set of 3,500 instructional image-text pairs. MiniGPT-4 relies on the integration of a pre-trained BLIP2 visual encoder(Li et al., 2023c) and a LLaMA language model.The model variant evaluated in our experiments is "MiniGPT4-aligned-with-llama2-7b". The evaluation process is completed over a period of 50 GPU hours, utilizing one

NVIDIA GeForce RTX 3090 GPU with 25GB of memory.

**LRV-instruct** aims to enhance the accuracy and robustness of multimodal artificial intelligence models when processing visual instructions. It trains models by including 120,000 positive and negative visual instructions to identify and avoid hallucinations that occur during task execution. The model variant evaluated in our experiments is "LRV-MiniGPT4-7b". The evaluation process is completed over a period of 60 GPU hours, utilizing one NVIDIA GeForce RTX 3090 GPU with 25GB of memory.

**Blip-2** is an efficient pre-training strategy that initiates visual language pre-training by utilizing pre-existing frozen pre-trained image encoders and frozen large-scale language models. The model variant evaluated in our experiments is "BLIP w/ ViT-L". The evaluation process is completed over a period of 40 GPU hours, utilizing one NVIDIA GeForce RTX 3090 GPU with 25GB of memory.

**Hallucidoctor** is a novel illusion detection and elimination framework based on the cross-checking paradigm, aiming to automatically identify and eliminate illusions in training data. The model variant evaluated in our experiments is "Minigpt4-vicuna-LLaVA+". The evaluation process is completed over a period of 80 GPU hours, utilizing one NVIDIA GeForce RTX 3090 GPU with 25GB of memory.

**GPT-4** is the latest generation of closed-source large language models released by OpenAI, which has seen significant enhancements in terms of the scale of data training and computational complexity. GPT-4V adds visual capabilities to GPT-4, enabling the model to not only understand and generate text but also to comprehend and analyze image content.The model variant evaluated in our experiments is "gpt-4-vision-preview". The evaluation process is completed over a period of 40 hours, using the API interface provided by OpenAI.

## D GQA dataset

The **GQA** (Hudson and Manning, 2019) dataset consists of real-world images with synthesized questions. Each image is associated with a cleaner scene graph of image objects, their attributes, and relations. Each question is associated with a structured functional program, that refers objects and relations to specify the reasoning route in the scene graph for the final answer. This dataset is distributed under license CC BY 4.0.

## E Intended use of VisDiaHalBench

The VisDiaHalBench is distributed for research purposes. The VisDiaHalBench is constructed based on images and scene graphs from the GQA dataset, thus does not contains any information that names or uniquely identifies individual people or offensive content same to the GQA.

## F Detailed experimental results

Table 12,13,14,15 presents the detailed experimental results from different perspectives.

## G Supplementary examples

Figures 4, 5 show supplementary examples from VisDiaHalBench.

As a question asker, could you please provide a question according to my request? You will receive a "question type" where "exist" represents asking whether there is a object; "queryrel" requires asking about the relative relationship between obj1 and obj2; "verifyRel" indicates asking whether these two objects have a certain relative relationship,(for example, is the book on the left of a yellow thing?) "queryRel" asks for the relative relationship between the two objects. "verifyTargetRel" stands for the query whether A and B have the relationship mentioned in the "last round".(for example, does this relationship(last round) of position also exist between the banana and the table?)."verifyattr" should ask whether obj possesses a certain attribute; "verifyTargetAttr" asks whether a certain object possesses a specific attribute mentioned in the "last round".(for example, Do the book have the same color with it(last round)?). And for "queryAttr," the answer should be the attribute of the queried object.(For example, what is the color of the object on the table?). Next, you will receive a "last round," for example, "logo." If provided, please use pronouns (such as "it," "her") to refer to it in your question. Then, you will receive a "prompt" where <obj> represents the objects that can appear in your question (including their relationships and properties), and "<answer triplet>" represents the information about the subject object of the question.(example: "tractors, <name>, tractors" represents the object as "tractors"."bike, <attribute>, navy" represents the object as "bike" and it has the attribute "navy". Attention, please make sure to include its attribute when you mention the obj.) The "<answer>" serves as the correct answer for your question.You should make the best use of all the information available. Now based on this information:[question type:question type,last round":{last round input},"prompt":prompt.] ,generate a question:

Table 9: Prompt for GPT-4 to generate questions with fixed answers.

As a keen questioner, you need to generate three questions related to the current issue and image that include potentially information in the format [Q1, Q2, Q3] (e.g., Is A present in the image? Where is the specific location of A? What is the color of A?).image:<Img><ImageHere></Img>, question: {question}, ###your proposed question:

Table 10: The prompt of Q-Probing to generate relevant questions.

Give the following image: <Img>ImageContent</Img>. You will be able to see the image once I provide it to you. Please answer my questions according to the dialogue.###Human: History:{historty_dialogue}, Image now:<Img><ImageHere></Img>, Question:{question}. (Before answering, please take a moment to consider these questions (do not need to answer): {Q1, Q2, Q3}) ###Assistant:

Table 11: The prompt of Q-Probing to answer a question given generate relevant questions.

| Model | Y/N | | Object | | Attribute | | Relation | |
|---|---|---|---|---|---|---|---|---|
| | EM | F1 | EM | F1 | EM | F1 | EM | F1 |
| Cheetah | 11.79 | 26.37 | 5.99 | 24.99 | 4.44 | 23.03 | 3.50 | 28.20 |
| LRV-instruct | 12.95 | 34.08 | 8.45 | 35.74 | 13.26 | 37.17 | 1.64 | 35.30 |
| MiniGPT4 | 20.77 | 37.09 | 4.63 | 30.09 | 8.65 | 34.68 | 3.72 | 30.21 |
| Q-Probing | 23.02 | 39.78 | 3.88 | 32.11 | 10.72 | 36.27 | 8.01 | 34.26 |
| LLaVA | 23.93 | 37.73 | 3.54 | 28.86 | 18.89 | 39.85 | 9.03 | 36.28 |
| GPT-4V | 32.33 | 46.61 | 22.52 | 50.15 | 29.16 | 50.44 | 25.21 | 49.37 |

Table 12: Average evaluation results under different answer types.

| Model | exist | | query object | | query attribute | | query relation | | verify attribute | | verify relation | | verify target attribute | | verify target relation | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | EM | F1 | EM | F1 | EM | F1 | EM | F1 | EM | F1 | EM | F1 | EM | F1 | EM | F1 |
| Cheetah | 25.31 | 29.02 | 5.99 | 24.99 | 4.44 | 23.02 | 3.50 | 28.20 | 23.31 | 29.15 | 12.52 | 27.48 | 11.33 | 25.75 | 0.07 | 22.37 |
| LRV-instruct | 27.26 | 31.62 | 8.44 | 35.74 | 13.25 | 37.17 | 1.64 | 35.29 | 16.10 | 26.89 | 13.57 | 37.87 | 14.31 | 37.23 | 0.01 | 34.85 |
| MiniGPT4 | 52.58 | 52.74 | 4.63 | 30.09 | 8.65 | 34.69 | 3.71 | 30.20 | 43.78 | 46.77 | 19.87 | 35.94 | 6.70 | 22.73 | 0.21 | 26.98 |
| Q-Probing | 59.41 | 59.57 | 3.87 | 32.11 | 10.72 | 36.27 | 8.00 | 34.27 | 53.38 | 55.98 | 21.36 | 36.84 | 9.90 | 23.20 | 1.87 | 27.68 |
| LLaVA | 49.56 | 49.56 | 3.54 | 28.86 | 18.88 | 39.85 | 0.92 | 36.27 | 38.59 | 40.15 | 20.29 | 30.10 | 36.72 | 45.46 | 0.07 | 16.95 |
| GPT-4V | 47.70 | 48.83 | 22.52 | 50.15 | 29.16 | 50.44 | 25.21 | 49.37 | 40.24 | 44.83 | 37.89 | 47.30 | 33.47 | 43.46 | 18.51 | 36.06 |

Table 13: Average evaluation results under different qeustion types.

| Model | None | | modify obj none exist | | modify attr none exist | | refer none exist | | remove none exist | |
|---|---|---|---|---|---|---|---|---|---|---|
| | EM | F1 | EM | F1 | EM | F1 | EM | F1 | EM | F1 |
| Cheetah | 19.50 | 27.86 | 0.01 | 19.45 | 0.45 | 25.31 | 0.22 | 23.27 | 0.21 | 24.50 |
| LRV-instruct | 21.55 | 33.76 | 0.01 | 33.04 | 0.01 | 31.72 | 0.01 | 35.48 | 0.01 | 33.24 |
| MiniGPT4 | 34.47 | 43.04 | 0.02 | 25.53 | 0.01 | 27.39 | 0.13 | 29.42 | 0.01 | 28.32 |
| Q-Probing | 36.82 | 46.51 | 0.03 | 22.54 | 4.13 | 30.03 | 1.65 | 28.26 | 0.97 | 29.15 |
| LLaVA | 39.78 | 48.26 | 0.01 | 18.15 | 0.15 | 26.97 | 0.06 | 25.92 | 0.01 | 0.25 |
| GPT-4V | 35.04 | 47.16 | 60.00 | 70.67 | 22.65 | 43.78 | 17.83 | 35.93 | 36.00 | 52.27 |

Table 14: Average evaluation results under different hallucination types.

| Model | remove | | modify attr | | modify obj | |
|---|---|---|---|---|---|---|
| | EM | F1 | EM | F1 | EM | F1 |
| Cheetah | 13.85 | 27.23 | 10.61 | 25.86 | 6.62 | 26.05 |
| LRV-instruct | 16.43 | 34.99 | 10.88 | 33.51 | 9.52 | 34.95 |
| MiniGPT4 | 26.68 | 40.04 | 17.37 | 35.38 | 7.62 | 30.82 |
| Q-Probing | 30.99 | 44.31 | 21.24 | 38.76 | 2.85 | 29.87 |
| LLaVA | 25.88 | 38.19 | 22.95 | 37.55 | 8.57 | 29.81 |
| GPT-4V | 40.12 | 50.59 | 31.04 | 45.94 | 16.67 | 39.39 |

Table 15: Average evaluation results under different edit types.

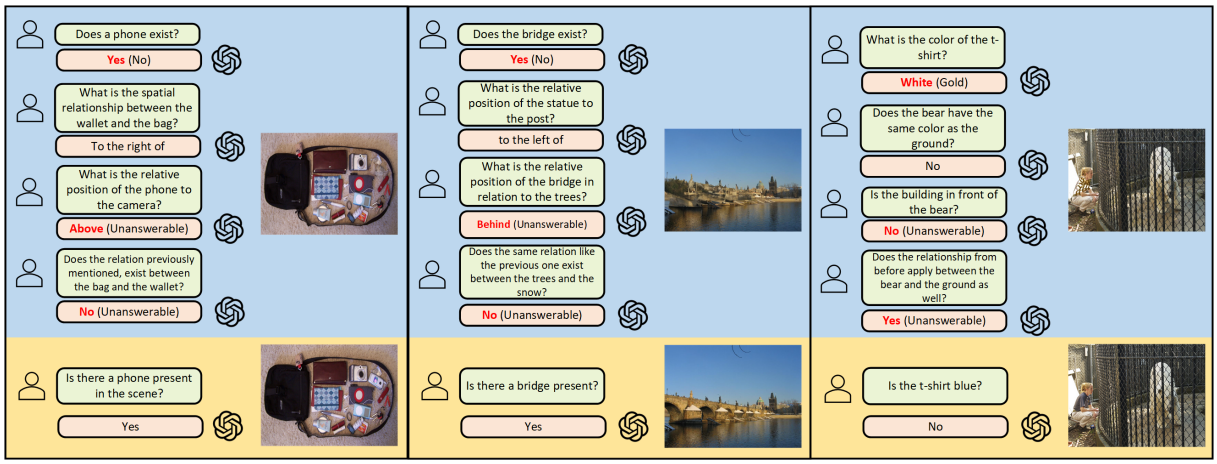


Figure 4: Supplement example 1 of VisDiaHalBench

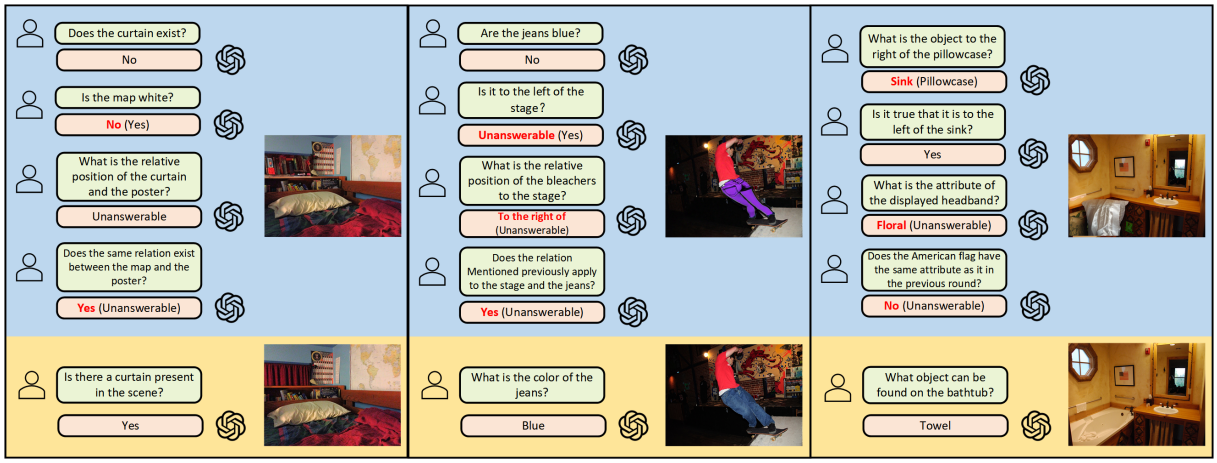


Figure 5: Supplement example 2 of VisDiaHalBench