Evaluating and Analyzing Relationship Hallucinations in Large Vision-Language Models

Mingrui Wu [1] **Jiayi Ji** [1] **Oucheng Huang** [1] **Jiale Li** [1] **Yuhang Wu** [1] **Xiaoshuai Sun** [1] **Rongrong Ji** [1]

Abstract

The issue of hallucinations is a prevalent concern in existing Large Vision-Language Models (LVLMs). Previous efforts have primarily focused on investigating object hallucinations, which can be easily alleviated by introducing object detectors. However, these efforts neglect hallucinations in inter-object relationships, which is essential for visual comprehension. In this work, we introduce R-Bench, a novel benchmark for evaluating Vision Relationship Hallucination. R-Bench features image-level questions that focus on the existence of relationships and instance-level questions that assess local visual comprehension. We identify three types of relationship co-occurrences that lead to hallucinations: relationship-relationship, subject-relationship, and relationship-object. The visual instruction tuning dataset's long-tail distribution significantly impacts LVLMs' understanding of visual relationships. Furthermore, our analysis reveals that current LVLMs tend to disregard visual content and overly rely on the common sense knowledge of Large Language Models. They also struggle with reasoning about spatial relationships based on contextual information.

1. Introduction

Recently, large language models (LLMs) such as GPT-4 (OpenAI, 2023) and Llama (Touvron et al., 2023) have demonstrated significant capabilities in addressing a broad spectrum of human-generated questions. The success of these models has spurred researchers to explore the use of LLMs in conjunction with visual inputs, leading to the development of various large vision-language models (LVLMs) (Li et al., 2023c; Liu et al., 2023c; Dai et al., 2023; Ye et al., 2023a). These endeavors typically involve methods like visual language pretraining (Li et al., 2023c) or visual instruction tuning (Liu et al., 2023c), aimed at integrating pre-trained visual encoders with LLMs to enhance their understanding of visual contexts.

However, despite their impressive performance, a significant challenge for these models is the unavoidable issue of hallucinations. Existing LVLMs often tend to generate responses that are inconsistent with the content of the images. This issue is particularly critical for LVLMs, which are expected to accurately comprehend images and produce answers consistent with the content of the visual input. While prior research has delved into evaluating object hallucinations (Li et al., 2023e), offering mitigation strategies through the object detection (Yin et al., 2023) or segmentation models (Wu et al., 2022; Chen et al., 2023c), there exists a notable gap in addressing hallucinations related to inter-object relationships, as shown in Figure 1. The latter, compared to object hallucinations, better reflects the LVLM's capacity to comprehend the intricacies of visual scenes. Currently, there is a shortage of comprehensive and rigorous benchmarks, to address these relationship hallucinations.

In this study, we introduce a novel Relationship Hallucination Benchmark (R-Bench) designed specifically for assessing relationship hallucinations in LVLMs. This benchmark comprises image-level and instance-level questions, labeled as 'Yes' or 'No', similar to the POPE evaluation (Li et al., 2023e). Image-level questions assess the existence of relationships in the image, while instance-level questions focus on specific object relationships, indicated by color bounding boxes or masks. The instance-level questions showcase the local visual understanding ability of LVLMs, adaptable to existing models without requiring retraining.

For the benchmark, we employ a combination of automatic generation by the Large Language Model (LLM) and manual curation. To ensure the benchmark's integrity, it is based on the nocaps validation set (Agrawal et al., 2019), preventing overlap with the pre-trained data of LVLMs. The construction process involves parsing all COCO captions to create a comprehensive relationship set. For each image in the nocaps dataset, we parse the provided captions

[1]Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, 361005, P.R. China. Correspondence to: Jiayi Ji <jjyxmu@gmail.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

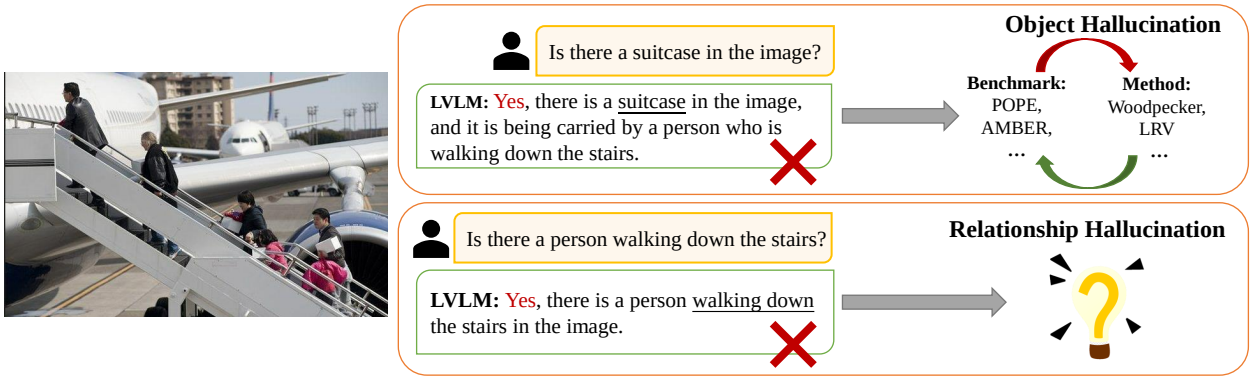


Figure 1. The object hallucination and relationship hallucination in Large Vision-Language Models. While substantial research has addressed object hallucinations in LVLMs, the issue of relationship hallucinations remains under-explored.

into relationship triplets, which are then aligned with the relationship set to establish a set of relationship seeds. Leveraging GroundingDINO, we identify significant objects with bounding boxes. Subsequently, we create prompts based on nocaps captions, relationship seeds, and bounding boxes, which are then fed into LLM to generate both image-level and instance-level questions. Finally, after rigorous manual selection, we have established R-Bench. The total number of questions retained after the filtering process amounts to 12,655, of which 7,883 are image-level questions, and 3,768 are instance-level questions.

We assess various recently popular LVLMs on our R-Bench and present our findings as follows: 1) Relationship hallucinations in LVLMs are more severe than object hallucinations, mainly due to the long-tail distribution between relationships and objects in the training data. 2) Relationship hallucinations often emerge from the co-occurrence patterns among relationships, specifically relationship-relationship, subject-relationship, and relationship-object. 3) Employing fine-grained image-text alignment could potentially mitigate hallucinations. Additionally, our analysis of counterfactual and illusion relationship hallucinations, based on web-collected images, reveals: 4) Existing LVLMs often overlook visual content, relying on LLM's common sense for predictions. 5) LVLMs struggle to reason about spatial relationships based on context. We aspire that our findings will stimulate the community to explore innovative solutions for mitigating relationship hallucinations in LVLMs.

2. Related Work

2.1. Large Vision-Language Model

Motivated by the achievements of Large Language Models (LLMs) (OpenAI, 2023; Touvron et al., 2023), there is a current trend among researchers to develop a diverse array of Large Vision-Language Models (LVLMs) (Li et al., 2023c; Liu et al., 2023c; Dai et al., 2023; Ye et al., 2023a; Li et al., 2023b; Chen et al., 2023a; Team, 2023; Du et al., 2022; Gao et al., 2023; Bai et al., 2023; Luo et al., 2023). These LVLMs typically consist of a visual encoder, a language decoder, and an image-text alignment module, in which the visual encoder and the language decoder are usually from pre-trained models, such as CLIP (Radford et al., 2021), Llama (Touvron et al., 2023) and Vicuna (Chiang et al., 2023). The image-text alignment module usually is trained on image-text pairs and finetuned through visual instruction tuning to achieve powerful visual conversation ability. However, the training data of LVLMs usually lean towards an intermediate understanding with a focus on individual objects, neglecting the crucial visual comprehension skill of inter-object relationships.

2.2. Hallucination in LVLMs

Presently, all LVLMs experience visual hallucination issues, often producing answers inconsistent with image content. Unlike LLMs, these visual hallucinations in LVLMs pose potential real-world harm, impacting their practical applications. There are lots of benchmarks (Lu et al., 2024; Yue et al., 2023; Kembhavi et al., 2016; Fu et al., 2023; Liu et al., 2023e; Li et al., 2023a; Yu et al., 2023; Wu et al., 2023a; Guan et al., 2023; Masry et al., 2022; Jing et al., 2023; Cheng et al., 2023; Villa et al., 2023; Wang et al., 2023b;a) for evaluating LVLMs. However, the hallucination benchmark is still under-explored. Previous research (Li et al., 2023e) has focused on exploring object hallucinations, which indicate the presence of objects in the expressions generated by LVLMs that either do not exist in the image or are inconsistent with the image content. Such hallucination can be alleviated by incorporating visual models, such as object detection models (Yin et al., 2023), segmentation models (Wu et al., 2022; Chen et al., 2023c) or other meth-

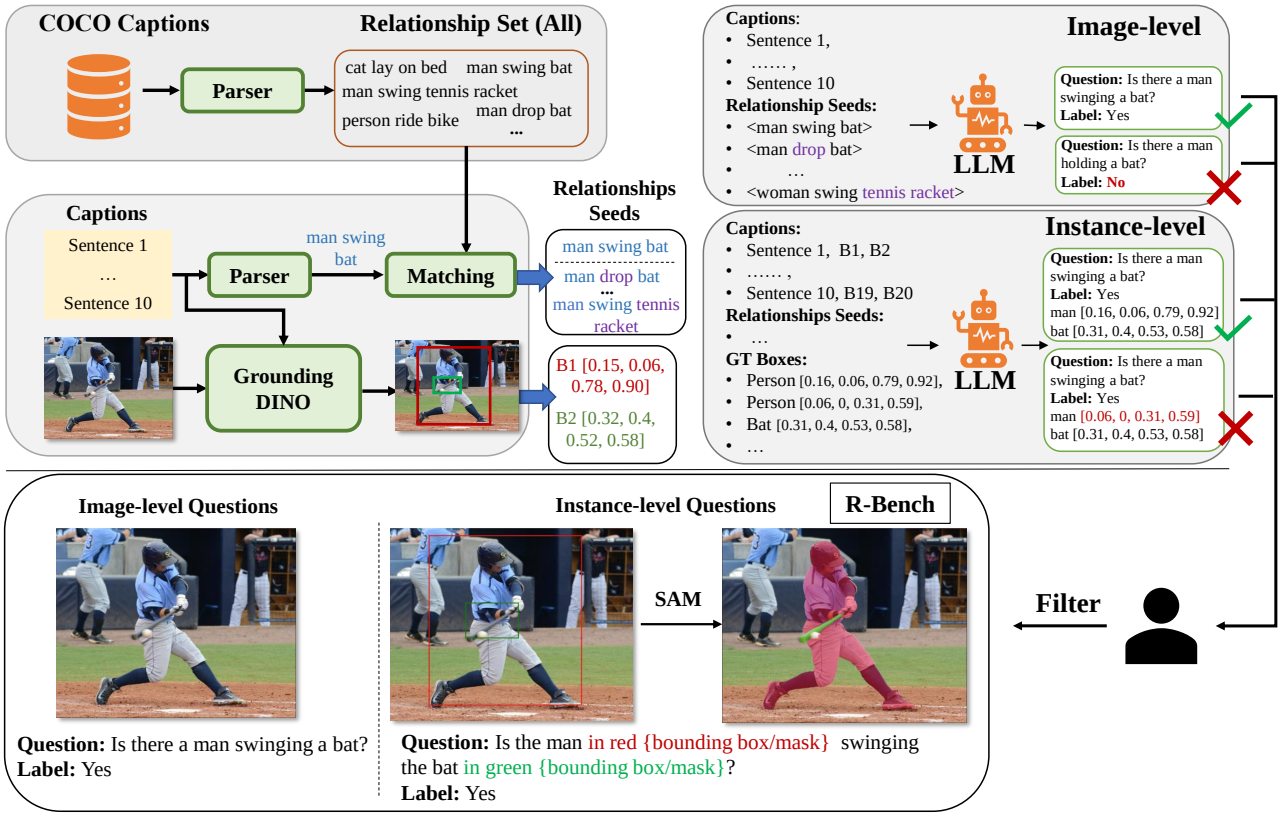


Figure 2. Our pipeline generates image-level and instance-level questions. First, we parse all COCO captions into a relationship set. Given a nocaps image, we parse its corresponding captions into relationship triplets and match these with the relationship set to obtain a set of relationship seeds. Using GroundingDINO, we identify significant objects with bounding boxes. We then create two types of prompts based on the nocaps captions, relationship seeds, and bounding boxes. Finally, we feed these prompts into an LLM to generate image-level and instance-level questions. Additionally, we carefully filter out noisy questions to create the refined R-Bench.

ods (Liu et al., 2023a;a; Li et al., 2023d; Chen et al., 2023b; Gunjal et al., 2023; Wang et al., 2024; Hu et al., 2023; Zhai et al., 2023; Zhou et al., 2023; Tang et al., 2023; Zhao et al., 2023; Leng et al., 2023; Lu et al., 2023; Zhang et al., 2021; Wu et al., 2023b; 2024). Some studies (Chen et al., 2023c; Villa et al., 2023) have begun to explore object attributes or relative positional relationships between objects hallucinations in LVLMs. Yet, there is a shortage of comprehensive and rigorous benchmarks and lacks effective analysis of hallucinations in inter-object relationships. This paper introduces a novel relationship hallucination benchmark using the nocaps validation set, conducting a thorough assessment to enhance LVLMs' image comprehension capabilities.

3. Relationship Hallucination Benchmark

In this section, we describe our Relationship Hallucination Benchmark (R-Bench). Although the pioneering efforts of benchmarks (Chen et al., 2023c; Wang et al., 2023a) delve into the relationship hallucination, yet inadvertently overlook potential issues related to data leakage. These benchmarks are typically constructed directly from the overlapped subset of COCO (Lin et al., 2014) and Visual Gnome (Krishna et al., 2017). However, it is noteworthy that the Visual Gnome dataset has been extensively employed in pre-training or visual instruction tuning across the majority of LVLMs (Liu et al., 2023b; Li et al., 2023c; Ye et al., 2023b). In light of this, we introduce a novel relationship hallucination benchmark, carefully constructed using data from the validation set of nocaps (Agrawal et al., 2019). Notably, nocaps primarily serves as a captioning evaluator for LVLMs (Dai et al., 2023; Li et al., 2023c).

3.1. Benchmark Construction

We construct two types of benchmarks, the image-level benchmark which focuses on the existence of relationships, and the instance-level benchmark which assesses local visual comprehension. The pipeline is shown in Figure 2.

Image-level Benchmark. We construct questions with a

Table 1. The reference object accuracies. We report box accuracy and mask accuracy respectively. The result is for reference only due to ignoring the different words with the same semantics, this can lead to lower accuracy.

MODEL	BOX ACC.	MASK ACC.
LLAVA-1.5	71.04	82.60
INSTRUTBLIP	72.45	77.29
MPLUG-OWL2	70.71	75.74
QWEN-VL	76.25	70.66

yes or no response, rendering our benchmark potentially compatible with the existing object hallucination benchmark (Li et al., 2023e). Additionally, this simple binary classification problem demonstrates better stability compared to instruction-based methods (Li et al., 2023e). We adopt an LLM to generate the questions based on a pre-extracted set of relationships from COCO captions. This approach allows us to analyze the hallucinations of LVLMs more effectively, as most LVLMs exhibit the same bias as the COCO caption data distribution due to their visual instruction tuning data being derived from COCO captions.

Specifically, we first apply a scene graph parser (Li et al., 2023f) to extract relationship triplets from COCO captions. Subsequently, leveraging the LLM, we generate 'Yes' or 'No' questions based on both nocaps captions and the pre-extracted set of relationships. To manage the prompt length effectively, we present only image-related relationships in the prompt for each image. This involves parsing nocaps captions into relationship triplets using the scene graph parser and forming relationship seeds by matching relationships in the set that overlap with any two elements of the current triplet. Additionally, we guide the LLM to concentrate on generating questions that emphasize relationships and filtering out nonsensical negative questions.

Instance-level Benchmark. The trend in recent LVLMs (Bai et al., 2023; Zhang et al., 2023) is increasing towards incorporating additional inputs, such as bounding boxes or masks, to enhance local comprehension capabilities. Additionally, prioritizing the identification of relationships between specified objects aligns more closely with practical demands. To address this, we introduce an instance-level benchmark to assess the relationship between reference objects, demanding LVLMs to possess local visual comprehension. This is achieved by attaching colored bounding boxes or masks to the objects mentioned in the question and visually representing these colored elements on the image.

We first verify that existing LVLMs can distinguish reference objects. To assess this ability without considering whether the images have been encountered by LVLMs, we utilize the COCO validation set along with corresponding

Table 2. The data statics of our benchmark after artificial filtering. The 'Pos' and 'Neg' denote the number of positive and negative questions. The 'Obj' and 'Rel' denote categories number of objects and relationship (Relationship between subject and object, rather than the relationship triplet) respectively. The total number of the 'Image', 'Obj', and 'Rel' minus the overlap of the ones between image-level and instance-level.

TYPE	QUESTION		IMAGE	OBJ	REL
	POS	NEG			
IMAGE	5,134	2,749	3,657	2,602	587
INSTANCE	2,896	872	2,645	2,317	511
TOTAL	8,030	3,621	4,034	3,514	791

instance segmentation annotations, incorporating bounding boxes or masks. Since the existing LVLMs cannot detect and identify small objects, we first use COCO captions and GroundingDINO (Liu et al., 2023d) to extract the significant object area and match it with the ground truth to obtain the accurate object position. Subsequently, we visualize the identified objects by drawing bounding boxes or masks in random red or green colors on the images. To prompt the LVLMs, we use questions like "What is the object in the {red/green} {bounding box/mask}?". The typical response is "The object in the {red/green} bounding box is a {object}", allowing us to compute accuracy by matching the label with this response.

It is important to note that this approach may yield lower accuracy due to the complexities introduced by open-set questions, such as potential variations in object names with correct meanings, such as "bike" and "bicycle" (illustrated in Figure 10). We evaluate several recent LVLMs, and the results are presented in Table 1 for reference. Notably, all LVLMs demonstrate relatively high accuracy, indicating their proficiency in discriminating reference objects. Moreover, most LVLMs exhibit higher accuracy in mask-based object discrimination compared to box-based methods, likely due to the more precise references provided by masks. Notably, Qwen-VL (Bai et al., 2023) performs box-level image-text alignment training, thus getting a better performance on box accuracy.

Next, we generate the questions with bounding boxes and masks. We first feed the nocaps captions into the GroundingDINO to get a set of object bounding boxes, which usually have a stronger correlation with the captions. Then we let LLM generate the yes or no questions based on nocaps captions paired with extracted object bounding boxes, related relationship set, and some ground-truth bounding boxes of the image sourced from OpenImage (as the nocaps images are derived from OpenImage (Kuznetsova et al., 2020)). The LLM outputs questions containing bounding boxes for the subjects and objects of relationships. We

Table 3. Results of LVLMs on our R-Bench. We compute average scores of 5 random subsets, and each subset has 1:1 pos-neg questions. The 'box' and 'mask' denote types of instance-level questions with bounding box and mask respectively.

TYPE	MODEL	ACCURACY	PRECISION	RECALL	F1 SCORE	YES
IMAGE-LEVEL(ALL)	LLAVA-1.5	71.23	64.27	96.89	77.28	76.12
	INSTRUTBLIP	69.31	62.76	96.45	76.04	77.60
	MPLUG-OWL2	73.66	67.60	91.84	77.88	68.60
	QWEN-VL	79.19	76.43	84.99	80.48	56.15
IMAGE-LEVEL(SUBSET)	LLAVA-1.5	71.30	64.31	97.11	77.38	76.32
	INSTRUCTBLIP	69.21	62.60	97.11	76.12	78.41
	MPLUG-OWL2	74.66	68.40	92.67	78.71	68.48
	QWEN-VL	79.51	76.38	85.48	80.67	55.98
BOX	LLAVA-1.5	53.15	51.71	95.53	67.10	92.37
	INSTRUTBLIP	51.95	51.14	87.39	64.52	85.44
	MPLUG-OWL2	53.90	52.38	85.67	65.01	81.77
	QWEN-VL	58.82	55.56	88.17	68.16	79.35
MASK	LLAVA-1.5	53.44	51.89	94.50	66.99	91.06
	INSTRUTBLIP	55.63	53.61	83.51	65.30	77.88
	MPLUG-OWL2	55.80	55.46	80.44	65.65	76.17
	QWEN-VL	59.84	57.65	74.15	64.87	64.31

further feed these bounding boxes into the SAM (Kirillov et al., 2023) to get the corresponding masks. Upon obtaining questions with bounding boxes and masks, we visualize them on the image by incorporating colored bounding boxes or masks. When evaluating LVLMs on this instance-level benchmark, we present images adorned with these visual elements, and the prompts are formatted as "Is there {subject} in the red {bounding box/mask} relationship {object} in the green {bounding box/mask} in the image?".

This method not only enhances the model's ability for fine-grained analysis but also facilitates a direct comparison and integration with prior research.

3.2. Data Statics

We obtain a total of 24,897 questions. Due to insufficient image information provided to the LLM, it tends to generate numerous noisy questions. So we employ filters to filter the generated questions. These filters screen out questions if they meet any of the following criteria: (1) the question contains a typographical error; (2) the question is not related to relationships; (3) the answer label is incorrect or cannot be determined; (4) there is a discrepancy between the object described in the question and the object highlighted by the box or mask, among others. Filtering out the noise and ensuring logical and label-correct questions takes approximately two weeks.

The total number of questions after filtering is 11,651, in which the number of image-level questions is 7,883 with 5,134 positive questions and 2,749 negative questions, the number of instance-level questions is 3,768 with 2,896 positive questions and 872 negative questions, as shown in Table 2. The instance-level with box and mask share the same questions. The total image number is 4,034. The category number of objects in questions is 3,514, of which 3,028 objects appear in COCO captions. The category number of relationships in questions is 791, of which 739 relationships appear in COCO captions. And the relationship types include actions, spatial relationships, relationship existence, and so on. These objects and relationships together form a total of 11,335 relationship triples in the questions, of which 4,941 appear in COCO captions. On average, each question has a unique relationship triplet. Due to the severe imbalance between positive and negative questions, we randomly select an equal number of both to create a subset for LVLMs evaluation.

4. Experiments

4.1. Experimental Setup

Datasets and Evaluation Metrics. Our benchmark is built upon the validation set of nocaps (Agrawal et al., 2019), which consists of 4500 images from OpenImage (Kuznetsova et al., 2020) and each image with 10 captions. These images can be divided into 648 in-domain images (only with COCO classes), 2938 near-domain images (with COCO and novel classes) and 914 out-domain images (only with novel classes). We follow the POPE (Li et al., 2023e) to use Accuracy, Precision, Recall, F1 score, and Yes ratio as the evaluation metrics.

Implementations. We evaluate several recently popular LVLMs: LLaVA-1.5 (Liu et al., 2023c;b), InstructBLIP (Dai et al., 2023), mPLUG-Owl2 (Ye et al., 2023b), Qwen-VL (Bai et al., 2023). All models are evaluated on 1 NVIDIA

Table 4. Comparing image-level relationship hallucination with image-level object hallucination. We apply POPE to get an object hallucination set on the validation set of the nocaps. We construct two corresponding object questions for each relationship question for POPE adversarial, popular and random setting, and report the mean results.

TYPE	MODEL	ACCURACY	PRECISION	RECALL	F1 SCORE	YES
IMAGE-OBJ	LLAVA-1.5	78.34	71.01	97.81	82.12	69.54
	INSTRUCTBLIP	78.77	72.12	96.72	82.36	67.94
	MPLUG-OWL2	77.11	71.24	92.36	80.31	65.25
	QWEN-VL	86.57	93.29	78.99	85.50	42.42
IMAGE-REL	LLAVA-1.5	71.23	64.27	96.89	77.28	76.12
	INSTRUCTBLIP	69.31	62.76	96.45	76.04	77.60
	MPLUG-OWL2	73.66	67.60	91.84	77.88	68.60
	QWEN-VL	79.19	76.43	84.99	80.48	56.15

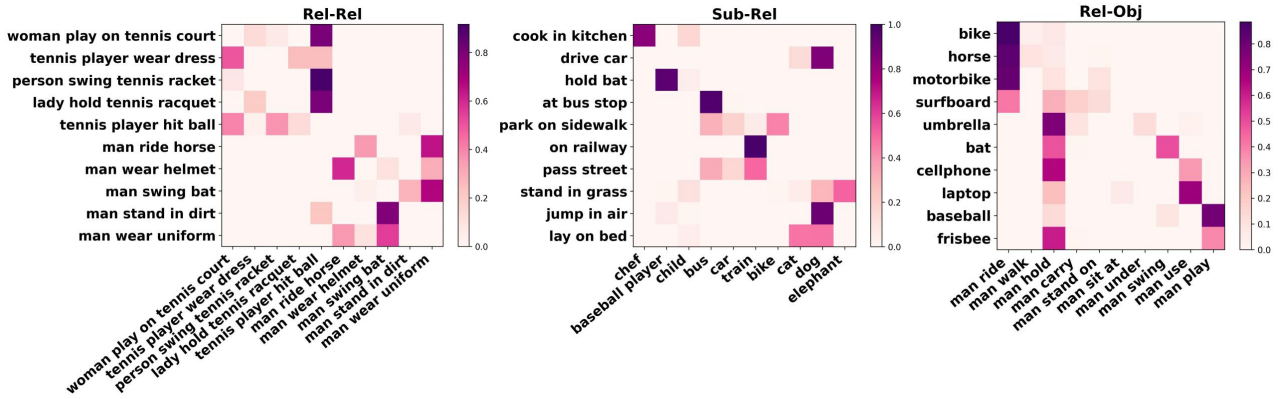


Figure 3. The co-occurrence matrices constructed between relationship-relationship (left), subject-relationship (middle), and relationship-object (right) respectively. The matrices show the conditional probability that an element of the y-axis occurs when another element of the x-axis is happening.

RTX-3090 with 24G memory. And the LLM used to generate questions is Llama2-chat-13B (Touvron et al., 2023). More information about models is shown in the Appendix.

4.2. Evaluating on the R-Bench

Evaluating LVLMs. We evaluate several popular LVLMs on our Relationship Hallucination Benchmark (R-Bench) under three settings: image-level, instance-level with bounding box, and instance-level with mask. For each setting, we randomly sample 5 subsets with a 1:1 positive-negative question ratio and compute the average scores of these 5 subsets. Additionally, we evaluate an extra image-level subset using the same set of images as the instance-level subsets, allowing us to more effectively assess the differences in hallucination issues between image-level and instance-level questions for LVLMs.

The results are shown in Table 3. In contrast to image-level hallucinations, the LVLMs consistently hallucinate on instance-level settings. Although these LVLMs have a good recognition ability for reference objects as shown in Table 1, they all fail to discriminate the relationship between reference objects. We show examples of image-level and instance-level relationship hallucinations in Figure 9, the LVLM can answer image-level questions well, but falls in instance-level questions. We believe that a fine-grained image-text alignment might help improve this.

Compare Relationship Hallucination with Object Hallucination. We assess the significance of addressing relationship hallucinations in LVLMs by comparing them with object hallucinations. To compare relationship hallucination with object hallucination, we use POPE to get an object hallucination set on the validation set of the nocaps. Specifically, we first perform POPE on the validation set of the nocaps based on the combined object labels of SEEM-based and Ground-Truth to obtain an object hallucination set. We construct two corresponding object questions for each relationship question for POPE adversarial, popular and random setting, and report the mean results.

The results are shown in Table 4. Compared with object

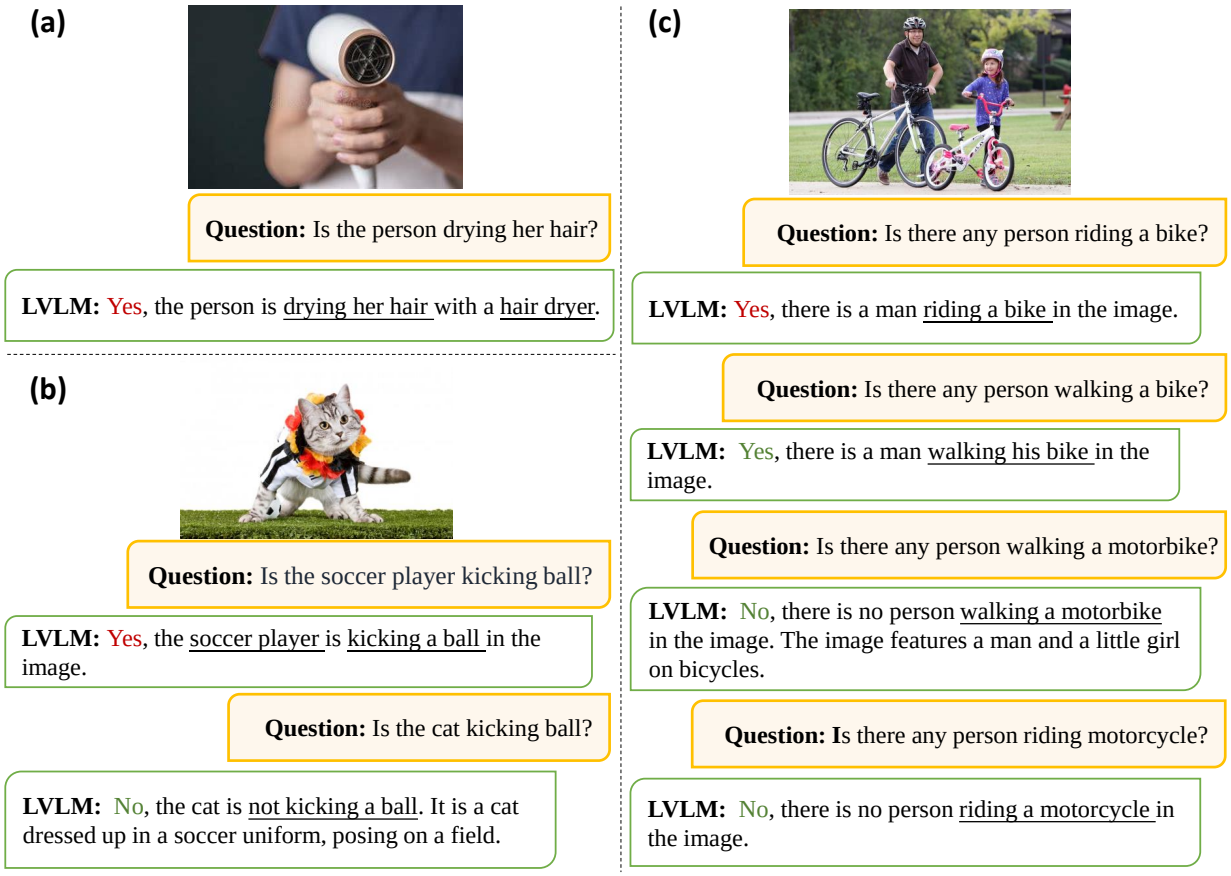


Figure 4. The examples of relationship hallucination which arise for different reasons. The wrong answers are marked in red, the relationships in the answers are underlined, and correct answers are marked in green.

hallucinations, existing LVLMs have more serious relationship hallucinations. Although Qwen-VL is an exception, it tends to answer no to ensure precision, which leads to a relatively low recall. Overall, existing LVLMs have more room for improvement in relationship hallucinations than object hallucinations.

4.3. Analysis of the Causes of Hallucinations

Analyzing the causes of relationship hallucinations will be more helpful for future work. Object co-occurrence is a major contributor to object hallucinations (Li et al., 2023e). Therefore, we analyze whether relationship co-occurrence also leads to relationship hallucination. We hypothesize three types of relationship co-occurrence that may lead to relationship hallucinations, including co-occurrence between the relationship-relationship, subject-relationship, and relationship-object. Unless otherwise specified, we perform the analysis on LLaVA-1.5.

Relationship-Relationship Co-occurrence. As the visual instruction tuning data is generated based on different captions, some related relationships will co-occur frequently, as shown in Figure 3 (left). Such as, when a "man swing bat", the "man stand in dirt" prone to appear at the same time. And we plot a bar chart in Figure 5 to show the probability of relationship hallucination when "man swings bat" occurs, the results show that frequently co-occurring relationships are more likely to result in hallucinations. And as shown in Figure 4 (a), we show an image, with a person holding a hair dryer and does not use it to dry the hair, to the LLaVA, however, it answers that "the person is drying her hair" incorrectly. So, the related co-occurring relationships may also contribute to hallucination.

Subject-Relationship Co-occurrence. There is usually a strong co-occurrence between the subject and certain behaviors in vision, for example, when the subject is a chef, he is usually cooking something instead of driving a car, as shown in Figure 3 (middle). This co-occurrence may lead to hallucinations, as shown in Figure 6 "bus" case. And as shown in Figure 4 (b), the image is a cat wearing a soccer

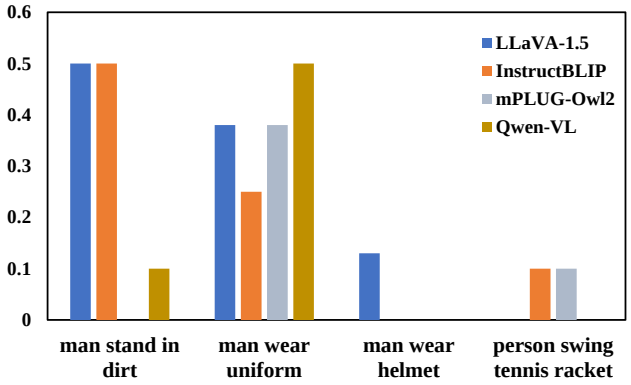


Figure 5. The probability of relationship hallucination when "man swings bat" occurs. The co-occurrence frequencies of these relationships with "man swings bat" decrease from left to right.

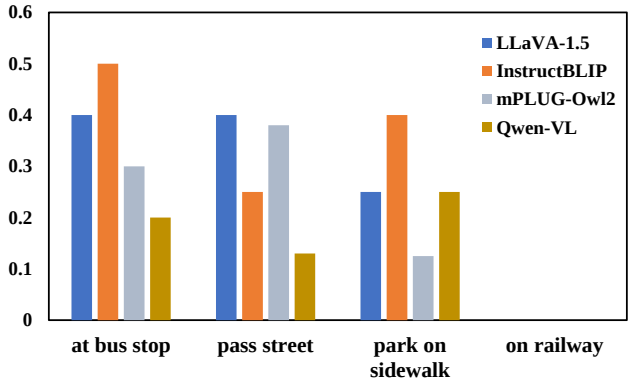


Figure 6. The probability of relationship hallucination when the subject is "bus". The co-occurrence frequencies of these relationships with "bus" decrease from left to right.

uniform with a soccer, when we ask LLaVA whether a "soccer player" kicking a ball, it answers "Yes" incorrectly. But when we ask it whether a "cat" kicking a ball, it can answer correctly. The relationship hallucinations resulting from this co-occurrence are widespread in LVLMs, due to the bias of the relevant training data towards the objects, as discussed in Relationship-Object Co-occurrence.

Relationship-Object Co-occurrence. There is also a co-occurrence between the relationship and the object, and some high-frequency co-occurrence relationships, such as riding a bike, also often lead to hallucinations, as shown in Figure 3 and Figure 7. And as shown in Figure 4 (c), it is hard for LVLM to distinguish between "riding bike" and "walking bike", but it is easy for it to distinguish whether the object the person is interacting with is a bike or a motorbike. This may be related to the unbalanced distribution of data in the caption data, as shown in Table 5 and Figure 11, the number of object words is much larger than that of relational words. Each caption contains an average of 3.6 object words, but only 1.3 relationship words. And 97,522 captions do not

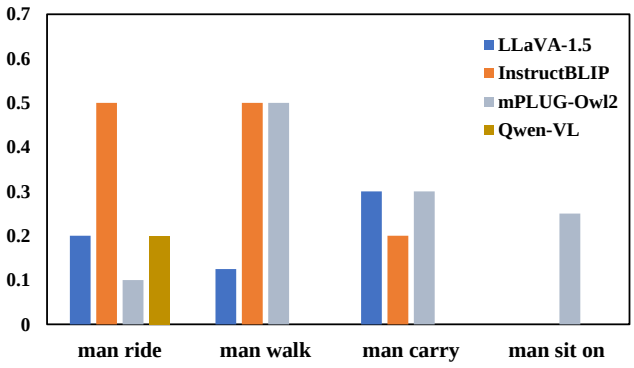


Figure 7. The probability of relationship hallucination when the object is "bike". The co-occurrence frequencies of these relationships with "bike" decrease from left to right.

Table 5. The data static of relationship words and object words in COCO captions. The 'Obj' and 'Rel' denote the number of objects and relationship words. The 'Spatial Rel' denotes the number of spatial relationship phrases whose proportion in relation words is about 1/5.

CAPTIONS	OBJ	REL	SPATIAL REL
591,753	2,130,794	764,711	159,649

contain relationship words, while only 419 captions do not contain object words. Thus forming a long-tail distribution between the relationship words and the object words.

4.4. Other Relationship Hallucinations

Most relationship hallucinations in R-Bench are linked to the three mentioned types of relationship co-occurrence, though some other relational hallucinations might be overlooked. Consequently, we further analyze additional relationship hallucinations in LVLMs, including counterfactual relationship hallucination and illusion relationship hallucination. We collect about 20 corresponding images from the web, and artificially form the questions with label no. We use accuracy as a metric to evaluate the LVLMs on these questions due to all the questions only with label no.

Counterfactual Relationship Hallucination. Existing LVLMs consist of a pre-trained visual encoder and LLM, and the LLM usually contains enough common-sense knowledge. Therefore, we assess whether LVLMs ignore the actual visual content and answer the question based on this common-sense knowledge directly. The result is shown in Table 6. The Qwen-VL has the best performance, yet only has an accuracy of 50%. As shown in Figure 8 (left), we show an image with a motorcycle consisting of only one wheel to the LVLM, but it misjudges that the motorcycle in the image has two wheels, even though this is true in common sense. This result shows that existing LVLMs tend to ignore the actual visual content and answer questions

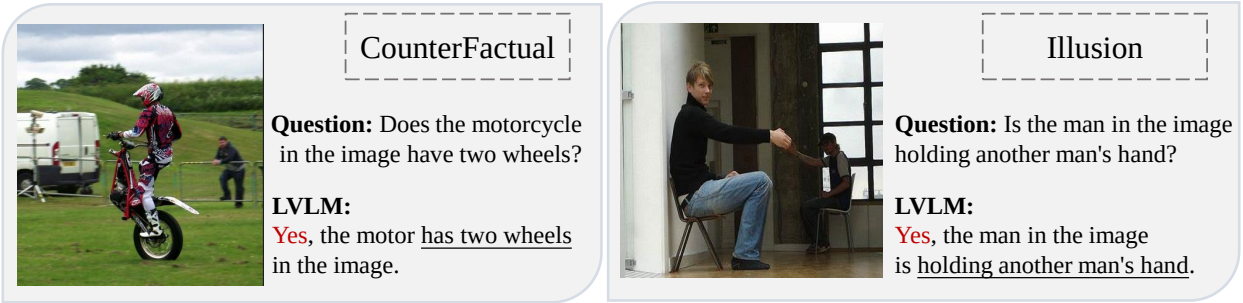


Figure 8. The examples of counterfactual relationship hallucination (left) and illusion relationship hallucination (right). The wrong answers are marked in red, and the relationships in the answers are underlined.

Table 6. Results of LVLMs on the other relationship hallucinations. We here report counterfactual accuracy, illusion accuracy, and total accuracy due to all questions with label no.

Model	Accuracy		
	Counter Factual	Illusion	Total
LLaVA-1.5	36.36	0.00	21.05
InstrutBLIP	9.09	25.00	15.79
mPLUG-Owl2	27.27	0.00	15.79
Qwen-VL	50.00	0.00	30.00

directly based on the knowledge from the LLM.

Illusion Relationship Hallucination. The ability to understand context-based spatial relationships demonstrates the reasoning ability of LVLMs, which can be evaluated through illusion images. The result is shown in Table 6 and Figure 8 (right). The best performance is 25% accuracy of InstructBLIP. The example as shown in Figure 8, there is a spatial dislocation here in the two men, and one can clearly tell this based on spatial relationships and context, but the LVLM argues that there is an interaction between these two spatially dislocated people. This indicates that the existing LVLMs still cannot reason about spatial relationships based on context.

5. Limitations

Firstly, our work exclusively focuses on analyzing relationship hallucinations in LVLMs, with plans to implement migration strategies based on our findings in future research. Secondly, our analysis covers only a subset of the causes behind relationship hallucinations. We anticipate future works to offer a more comprehensive and in-depth exploration of this phenomenon. Thirdly, a lack of an effective classification standard hinders a fine-grained categorization of relationship hallucination questions. Improved categorization could enhance the analysis of LVLM hallucinations. Fourthly, the manually filtered data may not be entirely free of noise due to omissions and biases in the filtering process. Lastly, we did not evaluate certain powerful closed-source LVLMs, like GPT-4, due to budget constraints.

6. Conclusion

In this paper, we are committed to evaluating and analyzing relationship hallucinations in LVLMs. We propose a novel Relationship Hallucination Benchmark (R-Bench) that includes image-level concentrate on relationship existence and instance-level questions for evaluating local visual comprehension. We analyze the relationship hallucination in the existing LVLMs on R-Bench and reveal several important relationship hallucination phenomena present in existing LVLMs. We anticipate that our findings will serve as inspiration for the community to delve into solutions addressing the relationship hallucinations of LVLMs.

Acknowledgements

This work was supported by National Key R&D Program of China (No.2022ZD0118201), the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No. U21B2037, No. U22B2051, No. 62072389), the National Natural Science Fund for Young Scholars of China (No. 62302411), China Postdoctoral Science Foundation (No. 2023M732948), the Natural Science Foundation of Fujian Province of China (No.2021J01002, No.2022J06001), and partially sponsored by CCF-NetEase ThunderFire Innovation Research Funding (NO. CCF-Netease 202301).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., and Anderson, P. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8948–8957, 2019.

Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., and Zhou, J. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., and Zhao, R. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023a.

Chen, Y., Sikka, K., Cogswell, M., Ji, H., and Divakaran, A. Dress: Instructing large vision-language models to align and interact with humans via natural language feedback. *arXiv preprint arXiv:2311.10081*, 2023b.

Chen, Z., Zhu, Y., Zhan, Y., Li, Z., Zhao, C., Wang, J., and Tang, M. Mitigating hallucination in visual language models with visual supervision. *arXiv preprint arXiv:2311.16479*, 2023c.

Cheng, S., Zhang, S., Wu, J., and Lan, M. Knvqa: A benchmark for evaluation knowledge-based vqa. *arXiv preprint arXiv:2311.12639*, 2023.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P., and Hoi, S. Instructblip: Towards general-purpose vision-language models with instruction tuning. arxiv 2023. *arXiv preprint arXiv:2305.06500*, 2023.

Du, Z., Qian, Y., Liu, X., Ding, M., Qiu, J., Yang, Z., and Tang, J. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 320–335, 2022.

Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., Wu, Y., and Ji, R. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Gao, P., Han, J., Zhang, R., Lin, Z., Geng, S., Zhou, A., Zhang, W., Lu, P., He, C., Yue, X., Li, H., and Qiao, Y. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.

Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., Manocha, D., and Zhou, T. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models, 2023.

Gunjal, A., Yin, J., and Bas, E. Detecting and preventing hallucinations in large vision language models. *arXiv preprint arXiv:2308.06394*, 2023.

Hu, H., Zhang, J., Zhao, M., and Sun, Z. Ciem: Contrastive instruction evaluation method for better instruction tuning. *arXiv preprint arXiv:2309.02301*, 2023.

Jing, L., Li, R., Chen, Y., Jia, M., and Du, X. Faithscore: Evaluating hallucinations in large vision-language models. *arXiv preprint arXiv:2311.01477*, 2023.

Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., and Farhadi, A. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pp. 235–251. Springer, 2016.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.

Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., and Bing, L. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *arXiv preprint arXiv:2311.16922*, 2023.

Li, B., Ge, Y., Ge, Y., Wang, G., Wang, R., Zhang, R., and Shan, Y. Seed-bench-2: Benchmarking multimodal large language models. *arXiv preprint arXiv:2311.17092*, 2023a.

Li, B., Zhang, Y., Chen, L., Wang, J., Yang, J., and Liu, Z. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023b.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023c.

Li, L., Xie, Z., Li, M., Chen, S., Wang, P., Chen, L., Yang, Y., Wang, B., and Kong, L. Silkie: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665*, 2023d.

Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., and Wen, J.-R. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023e.

Li, Z., Chai, Y., Zhuo, T. Y., Qu, L., Haffari, G., Li, F., Ji, D., and Tran, Q. H. FACTUAL: A benchmark for faithful and consistent textual scene graph parsing. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 6377–6390, Toronto, Canada, July 2023f. Association for Computational Linguistics. URL https://aclanthology.org/2023.findings-acl.398.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., and Wang, L. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 1(2):9, 2023a.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023b.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023c.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023d.

Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023e.

Lu, J., Gan, R., Zhang, D., Wu, X., Wu, Z., Sun, R., Zhang, J., Zhang, P., and Song, Y. Lyrics: Boosting fine-grained language-vision alignment and comprehension via semantic-aware visual objects. *arXiv preprint arXiv:2312.05278*, 2023.

Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.-W., Galley, M., and Gao, J. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*, 2024.

Luo, G., Zhou, Y., Ren, T., Chen, S., Sun, X., and Ji, R. Cheap and quick: Efficient vision-language instruction tuning for large language models. *arXiv preprint arXiv:2305.15023*, 2023.

Masry, A., Long, D., Tan, J. Q., Joty, S., and Hoque, E. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177.

OpenAI. Gpt-4 technical report, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Tang, L., Jiang, P.-T., Shen, Z., Zhang, H., Chen, J., and Li, B. Generalization and hallucination of large vision-language models through a camouflaged lens. *arXiv preprint arXiv:2311.11273*, 2023.

Team, I. Internlm: A multilingual language model with progressively enhanced capabilities. https://github.com/InternLM/InternLM, 2023.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Villa, A., Alcázar, J. C. L., Soto, A., and Ghanem, B. Behind the magic, merlim: Multi-modal evaluation benchmark for large image-language models. *arXiv preprint arXiv:2312.02219*, 2023.

Wang, J., Wang, Y., Xu, G., Zhang, J., Gu, Y., Jia, H., Yan, M., Zhang, J., and Sang, J. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023a.

Wang, J., Zhou, Y., Xu, G., Shi, P., Zhao, C., Xu, H., Ye, Q., Yan, M., Zhang, J., Zhu, J., et al. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*, 2023b.

Wang, L., He, J., Li, S., Liu, N., and Lim, E.-P. Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites. In *International Conference on Multimedia Modeling*, pp. 32–45. Springer, 2024.

Wu, H., Zhang, Z., Zhang, E., Chen, C., Liao, L., Wang, A., Li, C., Sun, W., Yan, Q., Zhai, G., et al. Q-bench: A benchmark for general-purpose foundation models on low-level vision. *arXiv preprint arXiv:2309.14181*, 2023a.

Wu, M., Zhang, X., Sun, X., Zhou, Y., Chen, C., Gu, J., Sun, X., and Ji, R. Difnet: Boosting visual information flow for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18020–18029, 2022.

Wu, M., Gu, J., Shen, Y., Lin, M., Chen, C., and Sun, X. End-to-end zero-shot hoi detection via vision and language knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2839–2846, 2023b.

Wu, M., Liu, Y., Ji, J., Sun, X., and Ji, R. Toward open-set human object interaction detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6066–6073, 2024.

Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023a.

Ye, Q., Xu, H., Ye, J., Yan, M., Liu, H., Qian, Q., Zhang, J., Huang, F., and Zhou, J. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*, 2023b.

Yin, S., Fu, C., Zhao, S., Xu, T., Wang, H., Sui, D., Shen, Y., Li, K., Sun, X., and Chen, E. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*, 2023.

Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., and Wang, L. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., Wei, C., Yu, B., Yuan, R., Sun, R., Yin, M., Zheng, B., Yang, Z., Liu, Y., Huang, W., Sun, H., Su, Y., and Chen, W. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2023.

Zhai, B., Yang, S., Xu, C., Shen, S., Keutzer, K., and Li, M. Halle-switch: Controlling object hallucination in large vision language models. *arXiv e-prints*, pp. arXiv–2310, 2023.

Zhang, S., Sun, P., Chen, S., Xiao, M., Shao, W., Zhang, W., Chen, K., and Luo, P. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023.

Zhang, X., Sun, X., Luo, Y., Ji, J., Zhou, Y., Wu, Y., Huang, F., and Ji, R. Rstnet: Captioning with adaptive attention on visual and non-visual words. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15465–15474, 2021.

Zhao, Z., Wang, B., Ouyang, L., Dong, X., Wang, J., and He, C. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*, 2023.

Zhou, Y., Cui, C., Yoon, J., Zhang, L., Deng, Z., Finn, C., Bansal, M., and Yao, H. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*, 2023.

A. Implementation Detail of LVLMs

We evaluated the latest versions of several popular LVLMs and chose the largest version model we could run as much as possible. During inference, we perform low-bit quantization on the LLM to avoid the out of GPU memory. We list the models we evaluated as well as the parameters of their core components in Table 7.

Table 7. The specific parameters of evaluted LVLMs.

MODEL	VISION ENCODER	PARAMETERS	LLM	PARAMETERS
LLAVA-1.5	VIT-L/14	0.4B	VICUNA-V1.5	13B
INSTRUTBLIP	VIT-G/14	2.0B	VICUNA	13B
MPLUG-OWL2	VIT-L/14	0.4B	LLAMA-2-CHAT	7B
QWEN-VL	VIT-G/14	2.0B	QWEN	7B

B. Additional Results of LVLMs on our R-Bench

Nocaps datasets include in-domain, near-domain and out-domain images. We provide additional results of LVLMs on our R-Bench under in-domain, near-domain and out-domain subsets. We randomly perform 1:1 positive and negative sampling for each subset 5 times, and average the results, as shown in Table 8.

C. Impact of Color Selection.

We select red and green as the base colors for marking because these colors are generally more distinguishable and easier for the model to recognize compared to other colors. Additionally, red and green tend to stand out more prominently within images, aiding in clearer object identification. To investigate the impact of color selection on model performance, we conducted experiments where we either swapped the red and green colors or replaced them with gold and pink, as shown in Ta. The results showed that interchanging red and green had only a minor effect on performance. However, substituting red and green with gold and pink led to various degrees of performance decline across different models, except for LLaVA, which demonstrated strong robustness to the change in colors.

D. Examples of Instance-level Relationship Hallucination

We compare the image-level and instance-level relationship hallucination. As shown in Figure 9, the LLaVA can better answer image-level questions than the instance-level ones.

E. Visualization Examples

There are some visualization examples, including the reference object recognition in Figure 10, the coco image with captions in Figure 11 and the other examples of counterfactual relationship hallucination and illusion relationship hallucination in Figure 12.

F. Data Static

We show some of the high-frequency relationships that appear in R-Bench in Figure 13. And the high-frequency relationship hallucination of the LVLMs on the image-level and instance-level questions in Figure 14.

Table 8. Results of LVLMs on our R-Bench. We compute average scores of 5 random subsets, and each subset has 1:1 pos-neg questions. The 'box' and 'mask' denote types of instance-level questions with bounding box and mask respectively.

TYPE	SUBSET	MODEL	ACCURACY	PRECISION	RECALL	F1 SCORE	YES
IMAGE	IN-DOMAIN	LLAVA-1.5	67.08	61.07	96.71	74.86	80.28
		INSTRUCTBLIP	67.36	61.02	98.63	75.39	81.94
		MPLUG-OWL2	68.06	62.85	90.41	74.15	72.92
		QWEN-VL	78.88	78.32	80.28	79.28	51.61
	NEAR-DOMAIN	LLAVA-1.5	69.9	63.37	97.25	76.73	78.33
		INSTRUCTBLIP	67.7	61.71	96.73	75.35	80.0
		MPLUG-OWL2	72.37	66.44	92.7	77.4	71.21
		QWEN-VL	77.91	74.33	85.77	79.64	58.13
	OUT-DOMAIN	LLAVA-1.5	66.2	60.39	97.35	74.54	81.92
		INSTRUCTBLIP	65.43	59.97	96.12	73.86	81.44
		MPLUG-OWL2	70.29	65.1	89.5	75.37	69.86
		QWEN-VL	75.5	72.11	83.55	77.41	58.24
BOX	IN-DOMAIN	LLAVA-1.5	51.18	50.6	98.43	66.84	97.25
		INSTRUCTBLIP	52.75	51.57	89.8	65.51	87.06
		MPLUG-OWL2	49.02	49.37	82.35	61.72	83.33
		QWEN-VL	56.47	53.89	89.41	67.24	82.94
	NEAR-DOMAIN	LLAVA-1.5	53.39	51.85	94.75	67.02	91.36
		INSTRUCTBLIP	51.68	50.97	87.85	64.51	86.17
		MPLUG-OWL2	55.06	53.11	86.39	65.78	81.33
		QWEN-VL	58.67	55.53	86.96	67.78	78.29
	OUT-DOMAIN	LLAVA-1.5	52.89	51.55	96.14	67.11	93.25
		INSTRUCTBLIP	52.89	51.7	87.71	65.05	84.82
		MPLUG-OWL2	56.39	53.91	88.07	66.88	81.69
		QWEN-VL	61.33	57.18	90.12	69.96	78.8
MASK	IN-DOMAIN	LLAVA-1.5	55.29	52.89	96.86	68.42	91.57
		INSTRUCTBLIP	55.29	53.15	89.02	66.55	83.73
		MPLUG-OWL2	53.73	52.34	81.96	63.87	78.24
		QWEN-VL	61.76	59.05	76.47	66.62	64.71
	NEAR-DOMAIN	LLAVA-1.5	54.46	52.46	95.00	67.60	90.54
		INSTRUCTBLIP	56.01	53.87	83.54	65.50	77.53
		MPLUG-OWL2	57.53	55.04	82.15	65.92	74.62
		QWEN-VL	58.39	56.38	74.05	64.01	65.66
	OUT-DOMAIN	LLAVA-1.5	54.64	52.54	96.02	67.92	91.39
		INSTRUCTBLIP	56.63	54.29	83.73	65.87	77.11
		MPLUG-OWL2	55.24	53.51	79.76	64.05	74.52
		QWEN-VL	63.49	60.33	78.8	68.33	65.3

Table 9. Impact of the color selection.

TYPE	MODEL	ACCURACY	PRECISION	RECALL	F1 SCORE	YES
BOX	LLAVA-1.5	53.15	51.71	95.53	67.10	92.37
	INSTRUCTBLIP	51.95	51.14	87.39	64.52	85.44
	MPLUG-OWL2	53.90	52.38	85.67	65.01	81.77
	QWEN-VL	58.82	55.56	88.17	68.16	79.35
BOX (SHIFT RED AND GREEN)	LLAVA-1.5	53.52	51.91	95.92	67.36	92.40
	INSTRUCTBLIP	52.57	51.51	87.48	64.84	84.91
	MPLUG-OWL2	53.80	52.40	82.71	64.16	78.91
	QWEN-VL	59.48	56.01	88.35	68.56	78.86
BOX (GOLD AND PINK)	LLAVA-1.5	53.38	51.81	96.79	67.49	93.41
	INSTRUCTBLIP	52.22	51.35	84.38	63.85	82.16
	MPLUG-OWL2	52.97	51.92	80.02	62.98	77.05
	QWEN-VL	57.97	55.14	85.55	67.06	77.58

Figure 9. The examples of comparing between the image-level (left) and instance-level (right) relationship hallucination. The wrong answers are marked in red, and the correct answers are marked in green.

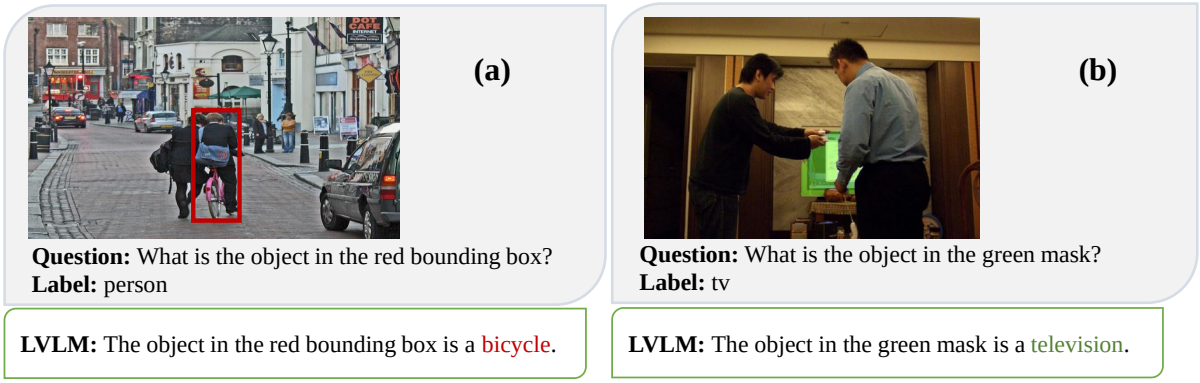


Figure 10. The examples of the reference object recognition base on bounding box (left) and mask (right). The wrong answers are marked in red, and the correct answers are marked in green. The mask provides more accurate reference than the box. And the LVLM generates the correct responses, with objects in different forms but with correct meanings.

Caption-1: A row of beach chairs with umbrella's above them on the beach.
Caption-2: A sandy beach with many chairs and umbrellas.
Caption-3: Some yellow chairs and yellow umbrellas on the beach.
Caption-4: A beach with water nearby and yellow umbrellas.

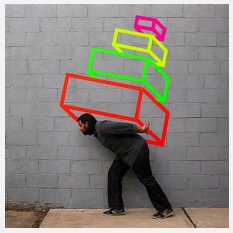
Caption-1: The photographer used a panoramic view for this photo.
Caption-2: Big living room with couches, TV, fireplace, and shelving.
Caption-3: There are multiple pieces of furniture in a living room.
Caption-4: A panoramic photo is shown of a room.

Figure 11. The examples of the coco image with captions. There are many image without any relationship descriptions or with unbalanced descriptions between the relationship and the objects.

CounterFactual

Question: Does the horse in the image have a tail?

LVLM: Yes, the horse in the image has a tail.



Illusion

Question: Is there person carrying box in the photo?

LVLM: Yes, there is a person carrying a box in the photo.

Figure 12. The other examples of counterfactual relationship hallucination and illusion relationship hallucination.

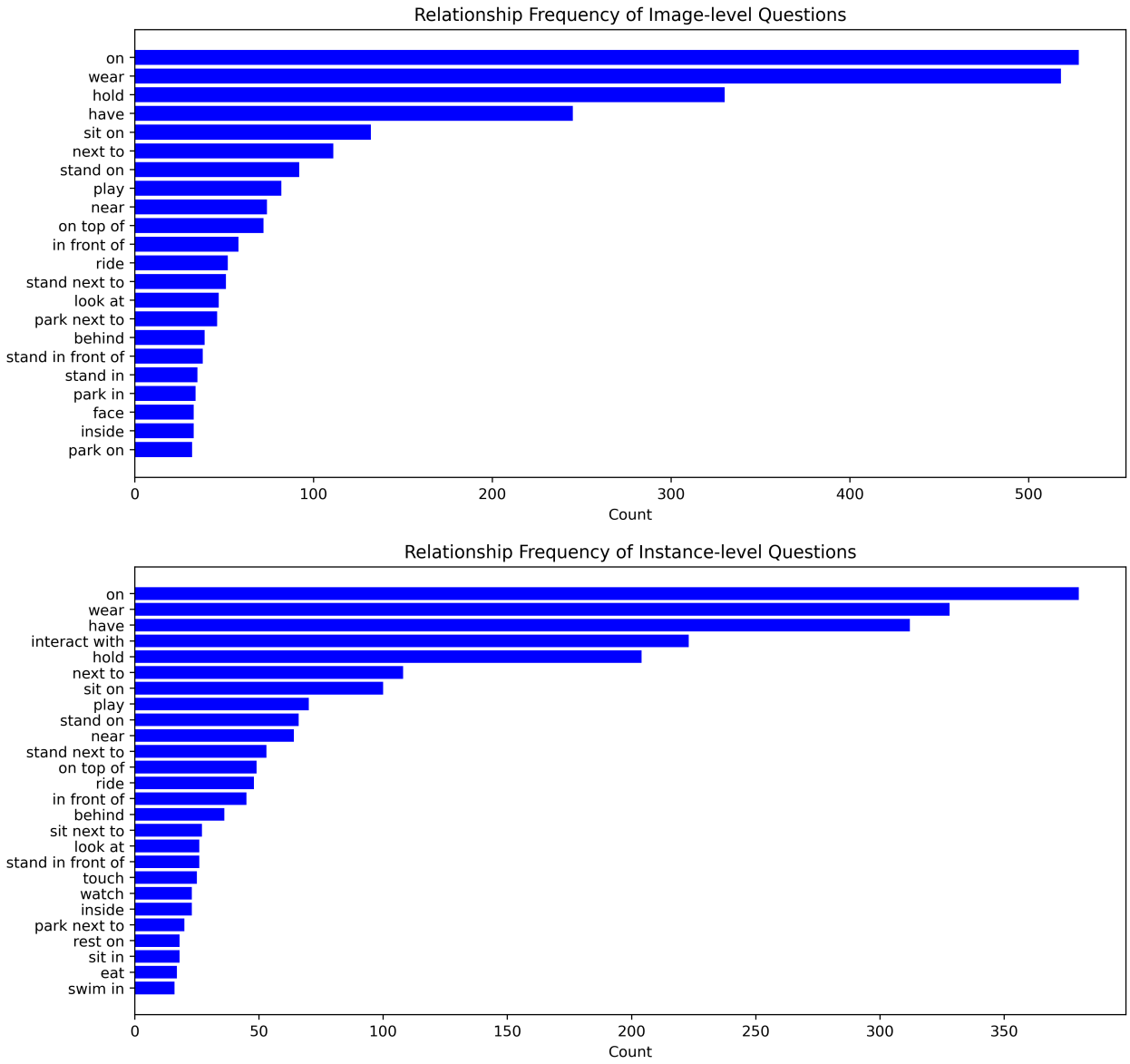


Figure 13. The relationship frequency in the image-level and instance-level questions.

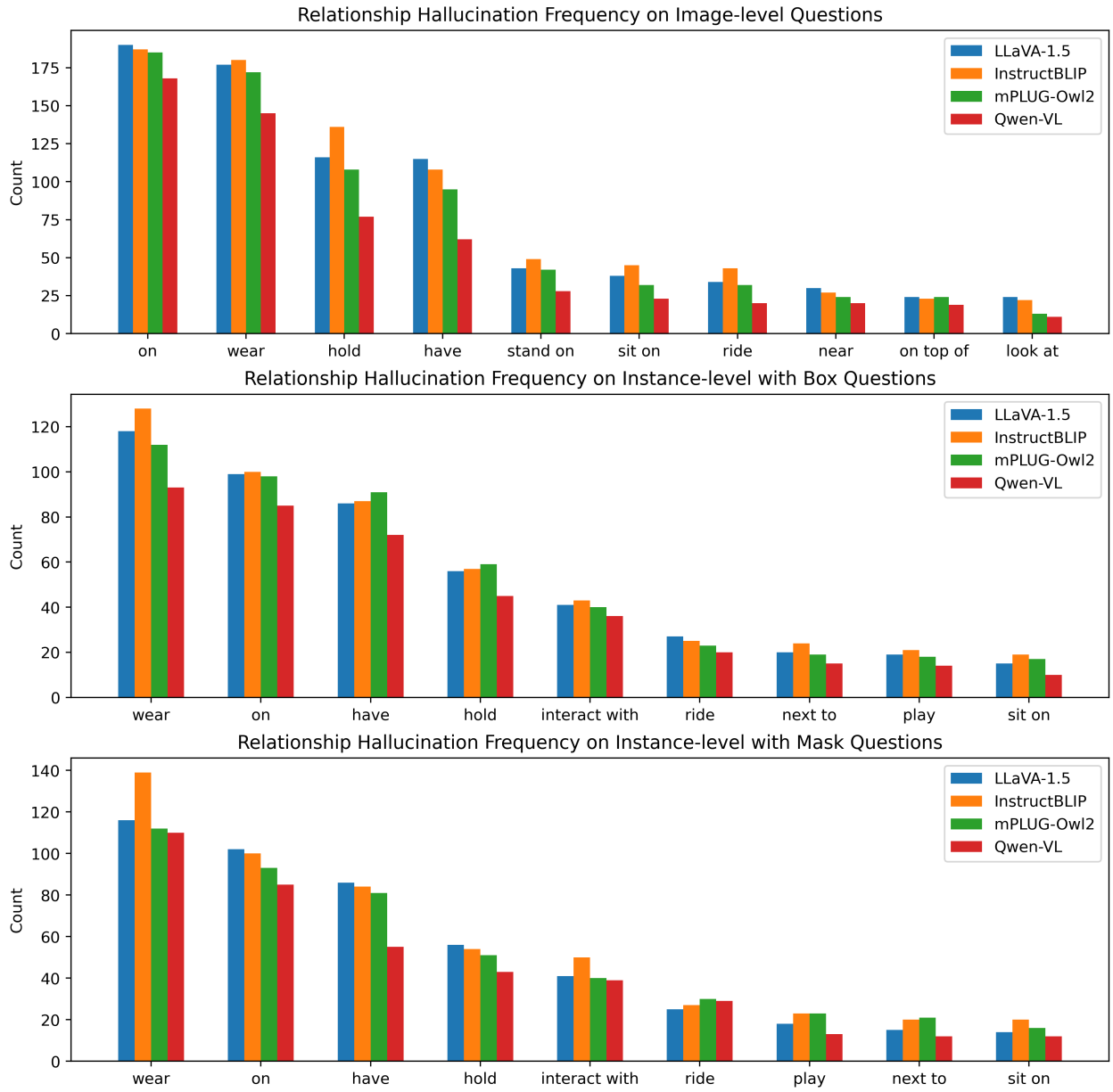


Figure 14. The relationship hallucination frequency of the LVLMs on the image-level and instance-level questions.