FIHA: Fine-grained Hallucinations Evaluations in Large Vision Language Models

Anonymous ACL submission

Abstract

The rapid development of Large Vision Language Models (LVLMs) often comes with widespread hallucination issues, making cost-effective and comprehensive assessments increasingly vital. Therefore, we introduce the FIHA (Fine-graIned Hallucination evAluation), a multidimensional hallucination evaluation method for LVLMs that is LLM-free and annotation-free. FIHA can generate QA pairs on any image dataset at minimal cost, enabling hallucination assessment from both image and caption. Based on this approach, we introduce a benchmark (FIHA-v1) consisting of diverse questions on various images from MS COCO and Foggy Cityscapes. Furthermore, we use the Davidson Scene Graph (DSG) to organize the structure among QA pairs, in which we can increase the reliability of the evaluation. We evaluate representative models using FIHA-v1, highlighting their limitations and challenges. Our code and data can be found here: anonymized link

1 Introduction

Large Vision-Language Models (LVLMs) such as MiniGPT-4 (Zhu et al., 2023) and LLaVA (Liu et al., 2023b), which extend Large Language Models (LLMs) by incorporating visual encoders, have shown prominent capabilities in visual understanding and generation (Zhang et al., 2024). However, LVLMs suffer from the issue of hallucination, which can lead to misinterpretation or erroneous assertions of the visual inputs, thus hindering the performance of models in multi-modal tasks (Huang et al., 2023). Specifically, the models may describe objects that do not exist in the image or incorrect object attributes and their relation. Generating such unreliable content will greatly reduce the model's credibility. Therefore, it is crucial to establish a benchmark for evaluating the hallucination level of LVLMs.

Previous studies (Li et al., 2023d; Wang et al., 2023b,a), as shown in Table 1, primarily employ a Question Generation (QG) module to create a set of validation questions and expected answers. These generated questions are then used to evaluate hallucinations in LVLMs. Despite the compelling success of the existing work, they still face two main challenges: (1) The existing work overlooks the dependency between different kinds of questions. For example, if the answer to "Is there a bike?" is no, dependent questions like "Is the bike yellow?" should be skipped. (2) Additionally, most prior work heavily relies on human annotations (Wang et al., 2023a) or LLMs (Li et al., 2023c) to generate QA pairs used in hallucination evaluation, which can be costly or labor-intensive.

To mitigate these limitations, we propose Fine-grained Hallucination Evaluation (FIHA), an automatic evaluation framework for fine-grained and diverse hallucinations in large-scale vision-language models. The framework takes either images or captions as input to generate QA pairs by extracting objects, attributes, and entity relations from the images or captions. Then, it creates QA pairs by incorporating multiple forms of questioning (e.g., "what," "who," "which," etc.) and allowing for free-form responses. To organize all the QA pairs into a tree-like structure, we introduce the Davidson Scene Graph (DSG) (Cho et al., 2023). With DSG, the response at each leaf node depends on the correctness of the root node answer, increasing the difficulty of model inference. Our QA pairs cover various types of questions, including misleading, narrative, and interrogative questions. This type of tree structure allows for a progressive deepening of questions, enabling a comprehensive evaluation of the model's understanding of the image.

We make the following key contributions through this work:

- To the best of our knowledge, FIHA is the first automated hallucination evaluation frame-

Table 1: Comparison with other benchmarks. Dis. denotes Discriminative and Gen. denotes Generative.

Evaluation Methods	Discriminative Hallucination			Task Type		Use DSG	LLM Free	Annotation Free
	Object	Attribute	Relation	Dis.	Gen.			
POPE (Li et al., 2023d)	✓	×	×	✓	×	×	✓	✓
NOPE (Lovenia et al., 2023)	✓	×	×	✓	×	×	×	✓
CIEM (Hu et al., 2023a)	✓	✓	×	✓	×	×	×	✓
AMBER (Wang et al., 2023a)	✓	✓	✓	✓	✓	×	✓	×
MHaluBench (Chen et al., 2024)	✓	✓	×	×	✓	×	×	×
FIHA (ours)	✓	✓	✓	✓	✓	✓	✓	✓

work that is LLM-free and annotation-free. This approach not only scales efficiently but also minimizes labor and associated costs.

- Based on FIHA, we generate a fine-grained evaluation benchmark FIHA-v1 that includes QA pairs evaluating various types of hallucinations and the semantic dependency relation organized by DSG.

- We evaluate and analyze several mainstream open-source and close-source LVLMs with FIHA-v1, providing valuable insights into their performance.

2 Method

In this section, we mainly introduce the FIHA (Fine-graIned Hallucination evAluation), an effective framework for evaluating fine-grained hallucination in LVLMs without the need for any manual annotations. FIHA extracts information from both the image and its caption. Given an image I and its caption C, we aim to query the LVLMs with the generated questions Q^I for the image and Q^C for the caption, in order to assess whether the responses are hallucinated. Compared to POPE, we have not only added detection at the attribute and relation levels, but also the QA pairs are no longer limited to yes/no type answers.

An overview of our method is provided in Figure 1. The subsequent parts will detail the procedures to produce the QA pairs.

2.1 Extract Information from Caption

Many datasets have corresponding captions for images, which highly summarize the information in the images. At the same time, natural language is more abstract compared to directly observable images, and it is also more prone to illusions. Therefore, we use the corresponding captions from the image dataset to extract objects, attributes, and relations. Given a caption C describing the image, our final objective is to produce sets of question-answer tuples $\{Q_i^C, A_i^C, L_i^C\}_{i=1}^N$, where Q_i^C represents a question, A_i^C denotes the gold answer, and L_i^C indicates the label (i.e., object, attribute, and relation) of the corresponding QA pair. Our main challenge lies in generating a diverse set of questions that cover all the information encapsulated within the caption. Unlike previous methodologies (Wang et al., 2023a; Li et al., 2023d; Lovenia et al., 2023; Hu et al., 2023a) which rely on human annotators, our proposed approach utilizes traditional NLP models and tools to extract effective information. The overall method is described by the following steps:

Extract Object and Attribute from Caption spaCy is a free and open-source natural language processing (NLP) library capable of performing a variety of complex NLP tasks. We primarily utilize its part-of-speech tagging feature to extract objects while simultaneously extracting their corresponding attributes, such as numerals, adjectives, and verbs. Here, we have obtained all the ground truth objects and their attributes: $G_{O,A}^C = \{O_1 : A_1, O_2 : A_2, .., O_n : A_n\}$, where n is the number of objects.

Extract Relation from Caption Stanford CoreNLP offers a suite of powerful natural language processing capabilities, enabling users to easily perform various linguistic analyses on text, which is also the reason we chose it for extracting relations. Here, we have obtained all the ground truth relations: $G_R^C = \{R_1 : playing, R_2 : standing\ on, .., R_m : is\ with\}$, where m is the number of relations.

2.2 Extract Information from Image

Because the information carried by the image itself is much greater than what is described in the caption, we consider extracting the necessary information directly from the image. Given a image I, our final objective is to produce sets of question-answer tuples $\{Q_i^I, A_i^I, L_i^I\}_{i=1}^N$, where Q_i^I repre-

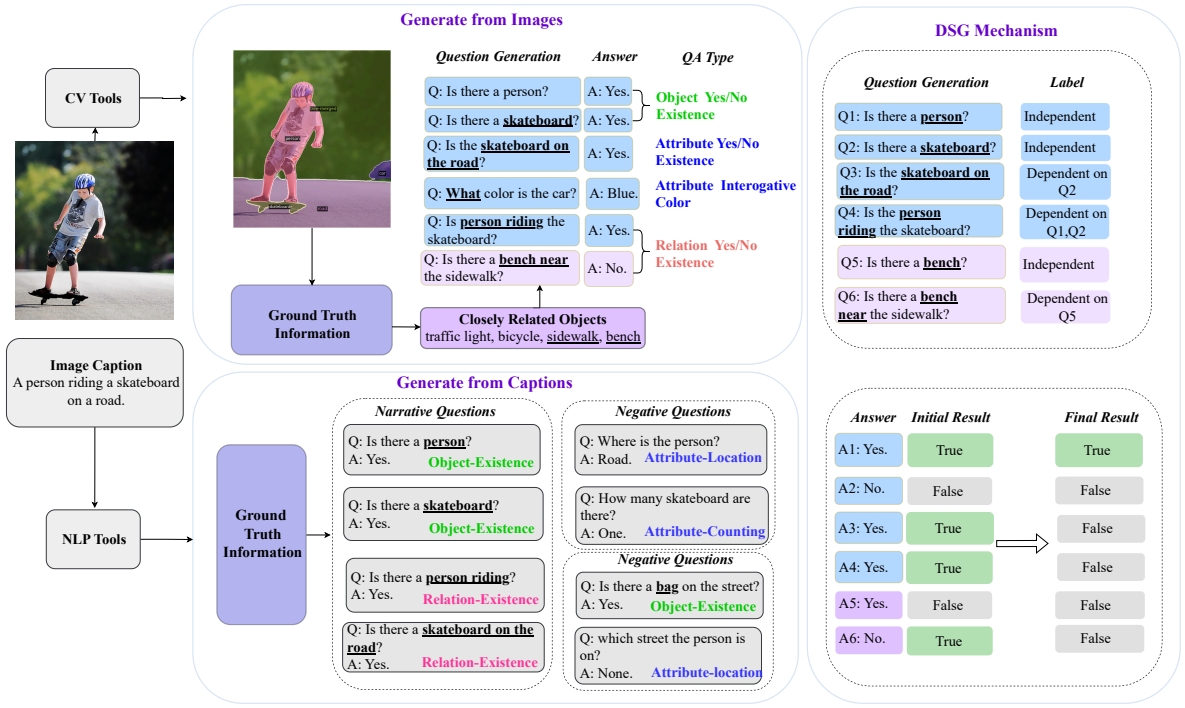


Figure 1: Overview of FIHA framework. FIHA extracts entities, attributes, and relations from images and captions respectively, and generates comprehensive and diverse questions to thoroughly detect model hallucinations. In the Figure, we can see that no LLM (Achiam et al., 2023) or additional manual annotations are used.

sents a question, A_i^I denotes the gold answer, and L_i^I indicates the label of the corresponding QA pair.

Extract Object and Attribute from Image Fast R-CNN (Girshick, 2015) is a fast object detection method based on Region-based Convolutional Networks, and it is a classic approach in the field of object detection. In this step, we obtain the ground truth objects and attributes.: $G^I_{O,A} = \{O_1 : A_1, O_2 : A_2, .., O_n : A_n\}$, where n is the number of objects.

Extract Relation from Image Here, we employ RelTR (Cong et al., 2022), a method for generating sparse scene graphs by decoding visual appearances and learning subject and object queries from the data. Using this method, we have obtained all the ground truth relations: $G^I_R = \{R_1 : behind, R_2 : near, .., R_m : wearing\}$, where m is the number of relations.

2.3 Generate Question Answer Pairs

At this point, we have obtained the objects, attributes, and relations as illustrated in Figure 2. Subsequently, questions will be formulated for querying LVLMs based on this information.

Question Formulation After obtaining the relevant information, we generate a series of questions

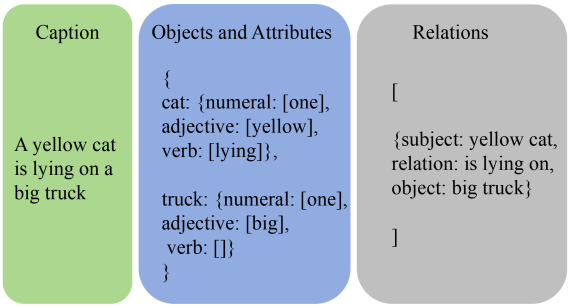


Figure 2: Example of extracted information.

that are directed at the object, attribute, and relation level. For the object level, we generate questions like "`Is there any {obj`$_k$`}?`" where {obj$_k$} comes from GO_i and CO_i. Similarly, we formulate diverse questions involving the attributes of objects such as "`What {color} is the {obj`$_k$`}?`" as well as relations between the objects such as "`Is there a {obj`$_1$`} near the {obj`$_k$`}?`". Similar to the method described in subsection 2.1 our formulated questions also include the interrogative and negative questions. To generate such a free-form formulation of questions, we prompt an LLM with some in-context examples so that meaningful questions are produced. After generating the questions we classify them into three

main and several subcategories similar as described in subsection 2.1.

Interrogative Questions In contrast to previous studies (Li et al., 2023d; Wang et al., 2023a) that focus solely on Yes/No questions for hallucination evaluation, our approach introduces greater diversity by generating questions incorporating interrogative words such as *"what," "who," "which," "where," and "how many"*. These questions elicit free-form responses, with no more than three words.

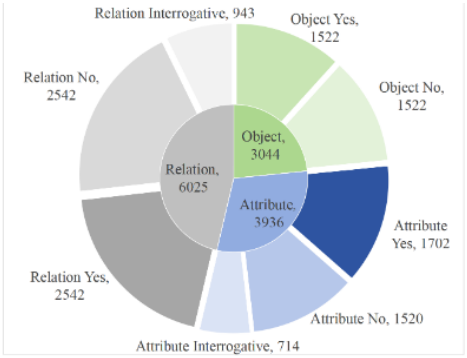
Negative Questions Motivated from Lovenia et al. (2023) we also focus on producing questions that elicit responses indicating the absence of objects, their attributes, and relations. Such questions are answered with negative pronouns such as *"none", "nobody", "nowhere", "zero", "neither"*, etc.

Misleading Questions According to Li et al. (2023d) LVLMs are prone to hallucinate the objects that mostly appear with the actual objects present in an image. Inspired by these insights, we have identified similar information from objects, attributes, and relations based on the ground truth, thereby generating misleading questions.

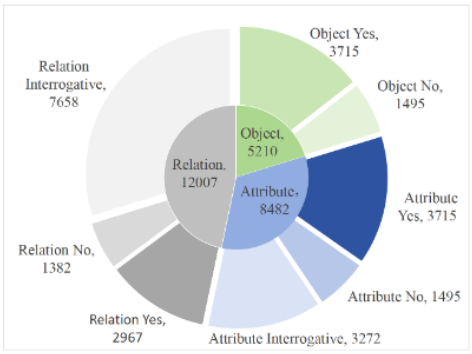
Narrative Questions At first, we generate questions that sequentially query a scene description, we denoted them narrative questions. In this framework, the Davidsonian Scene Graph (DSG) (Cho et al., 2023) plays a role similar to post-processing by chaining the QA pairs to form a tree-like structure. The root node addresses the existence of objects, followed by questions about the attributes of the root node objects in the next layer, and subsequently, questions about the relations between objects in the following layers. For instance, we have a list of questions $L^Q = \{Q_1 : Independent, Q_2 : Depends\ on\ Q_1\}$. Before determining if the answer to Q_2 is correct, we first assess Q_1, which concerns the accuracy related to the root node.

2.4 Answer Generation

In the case where the dataset lacks captions, this step is responsible for finding the answers to the formulated questions. We employ a pretrained VQA model (Li et al., 2023b) which provides answers to the formulated questions conditioned on the image. The findings by Li et al. (2023d) suggest that the small vision language models produce shorter answers with fewer hallucinations compared to mainstream LVLMs and therefore a reasonable choice for our task. The retrieved answer A_i to the question Q_i can have either Yes/No or free-form answers with no more than three words.



From Caption



From Image

Figure 3: Statistics of all issues generated by image and caption from MS COCO.

Table 2: The number of QA pairs generated from different datasets.

Dataset	From Image	From Caption
MS COCO (500)	25699	13007
Foggy Cityscapes (150)	7232	2801

3 Experiments

In this section, we randomly selected a subset of data from MS COCO and Foggy Cityscapes., used FIHA to generate corresponding QA pairs, and tested and analyzed the hallucination levels of some mainstream LVLMs.

3.1 Data Processing and Analysis

We randomly selected 500 images from the MS COCO dataset and 150 images from the Foggy Cityscapes. Using the process described in Section 3, we generated tens of thousands of QA pairs. The detailed data can be found in Table 2. Next, we will analyze the types of questions generated by the

Table 3: Evaluation results of LVLMs on questions generated from images and captions, respectively. F1 (Gen) refers to a text similarity metric, for which we employ BERTScore to measure the quality of responses to open-ended questions.

Model	Question Generated from Image					Question Generated from Caption				
	ACC	P.	R.	F1	F1 (Gen)	ACC	P.	R.	F1	F1 (Gen)
MS COCO										
mPLUG-Owl (Ye et al., 2023)	42.1	70.2	61.4	43.7	15.2	31.4	61.6	55.5	31.2	11.4
MiniGPT-4 (Zhu et al., 2023)	23.5	27.5	22.2	22.1	21.6	15.9	25.7	28.8	14.2	18.4
MultiModal-GPT (Gong et al., 2023)	59.1	46.4	47.1	46.6	16.1	23.8	39.6	45.7	22.1	10.8
LLaVA-1.5-7B (Liu et al., 2023b)	77.8	77.0	65.9	67.7	21.4	50.7	64.9	67.5	50.5	13.7
LLaVA-1.5-13B (Liu et al., 2023b)	78.9	80.9	66.4	68.3	20.9	47.6	64.2	65.5	48.5	13.8
InstructBLIP (Dai et al., 2023)	84.7	83.3	78.6	80.4	21.8	65.7	69.5	77.4	64.2	14.1
GPT-4V (OpenAI, 2024)	87.2	81.4	86.3	85.5	25.2	70.3	71.5	75.8	69.3	22.7
Foggy Cityscapes										
mPLUG-Owl (Ye et al., 2023)	64.8	60.2	51.1	42.7	18.6	29.5	58.9	51.6	25.6	29.3
MiniGPT-4 (Zhu et al., 2023)	30.1	30.2	27.6	28.1	9.4	23.4	34.4	37.8	23.0	11.6
MultiModal-GPT (Gong et al., 2023)	50.2	48.7	46.1	45.8	17.6	28.1	43.9	47.9	25.4	24.5
LLaVA-1.5-7B (Liu et al., 2023b)	67.7	68.4	56.2	52.9	19.7	29.1	50.0	49.2	25.8	27.5
LLaVA-1.5-13B (Liu et al., 2023b)	68.1	71.5	56.1	52.3	18.8	28.9	49.2	49.8	25.5	27.7
InstructBLIP (Dai et al., 2023)	70.9	75.6	60.2	58.8	20.3	32.8	58.3	53.2	30.5	29.2
GPT-4V (OpenAI, 2024)	76.3	70.1	64.6	66.0	16.2	33.7	53.3	51.7	32.1	21.7

pipeline.

The Figure 3 illustrates the distribution of question types generated from images and captions. The proportion of questions related to object, attribute, and relation is relatively balanced, reflecting the rationality of the method design. It is noteworthy that the abundance of the Interrogative category reflects FIHA's effective capability in generating tasks of the generation type, thereby enabling a more effective assessment of hallucinations.

3.2 Experimental Results

3.2.1 FIHA Overall Results

As shown in Table 3, the hallucination levels of the seven mainstream LVLMs evaluated using FIHA are presented. It's worth highlighting that GPT-4V excels in both image and caption QA pairs, achieving the best performance among the evaluated models. The second-best performer is Instruct BLIP, which significantly outperforms other models except GPT 4V across most metrics. Additionally, we have observed that model parameters are also significant factors affecting performance. For instance, LLaVA 13B provides a more comprehensive improvement over the 7B version.

3.2.2 FIHA Fine-Grained Results

Benefiting from the comprehensiveness of FIHA, we are able to evaluate the model's performance from more dimensions. Referencing the results of the Table 4, we will proceed with further analysis.

Object Hallucination It can be observed that even after introducing more negative samples, the *Accuracy* and *Precision* of the models remain high, indicating that most models have a strong capability to determine whether an object exists or not. In comparison, the *Recall* is somewhat lower, indicating that the model still has a tendency to lean towards affirmative responses.

Attribute Hallucination It is evident that this part of the hallucination is much more difficult to identify. Compared to the object itself, its color, quantity, size, and so on are indeed more challenging to judge. Even the best-performing GPT-4V has an F1 score of less than 80 on regular data. Moreover, the performance of the vast majority of models plummets on special datasets, indicating that the robustness of existing LVLMs needs to be enhanced.

Relation Hallucination This part is the most challenging, with GPT 4V's F1 score on regular data not even reaching 60. The DSG we introduced has further increased the difficulty; to accurately determine the answer to the relation within the tree structure, one must first correctly ascertain the existence of each of the two objects individually.

Table 4: The results of a more fine-grained assessment of LVLMs from the perspectives of object, attribute, and relation. The results are based on statistics from QA pairs generated by captions.

Model	Object				Attribute				Relation			
	ACC	P.	R.	F1	ACC	P.	R.	F1	ACC	P.	R.	F1
MS COCO												
mPLUG-Owl (Ye et al., 2023)	57.3	75.7	47.3	48.0	20.6	55.7	53.5	20.4	22.7	56.5	55.8	22.7
MiniGPT-4 (Ye et al., 2023)	66.2	59.5	62.6	59.5	9.6	12.8	9.2	9.4	4.7	12.1	11.4	4.9
MultiModal-GPT (Ye et al., 2023)	51.6	54.1	51.5	42.5	16.0	39.2	42.8	15.8	12.1	30.8	39.6	11.8
LLaVA-1.5-7B (Ye et al., 2023)	79.2	82.4	77.5	78.4	27.9	55.6	56.7	27.8	47.9	59.1	69.7	44.7
LLaVA-1.5-13B (Liu et al., 2023b)	70.8	80.6	70.2	68.3	34.3	56.4	59.7	33.7	42.1	58.3	66.6	48.1
InstructBLIP (Dai et al., 2023)	84.6	87.7	81.4	84.2	61.0	62.2	76.2	55.6	57.5	61.0	75.7	52.1
GPT-4V (OpenAI, 2024)	90.8	87.7	89.8	88.6	83.6	77.7	85.2	79.8	66.2	61.2	73.2	58.3
Foggy Cityscapes												
mPLUG-Owl (Ye et al., 2023)	52.9	32.3	50.0	39.2	15.7	54.8	52.1	15.3	11.8	34.6	46.9	11.1
MiniGPT-4 (Ye et al., 2023)	62.1	60.6	58.4	57.8	9.6	25.1	14.6	9.3	8.5	23.2	26.5	8.5
MultiModal-GPT (Ye et al., 2023)	52.9	59.7	52.6	42.1	12.6	33.9	38.6	12.5	11.5	33.3	39.4	11.4
LLaVA-1.5-7B (Ye et al., 2023)	54.0	63.3	54.0	44.4	11.5	33.2	46.0	10.8	15.4	47.8	48.9	15.1
LLaVA-1.5-13B (Liu et al., 2023b)	54.2	62.8	54.2	44.6	11.3	31.4	46.3	10.6	14.9	47.0	48.6	14.6
InstructBLIP (Dai et al., 2023)	54.2	65.2	53.9	44.2	20.7	55.1	54.6	20.6	15.9	48.5	49.2	15.6
GPT-4V (OpenAI, 2024)	61.8	69.6	59.2	54.5	11.1	37.0	33.1	11.0	20.4	50.5	50.4	20.3

Table 5: The performance decrease in various model metrics after introducing DSG.

Model	ACC↓	P.↓	R.↓	F1↓	F1 (Gen)↓
mPLUG-Owl (Ye et al., 2023)	29.6	22.1	14.0	28.7	14.2
MiniGPT-4 (Zhu et al., 2023)	62.6	51.8	62.1	61.2	42.3
MultiModal-GPT (Gong et al., 2023)	21.3	27.6	21.9	24.3	12.9
LLaVA-1.5-7B (Liu et al., 2023b)	4.2	11.7	4.5	4.8	5.7
LLaVA-1.5-13B (Liu et al., 2023b)	2.7	8.1	3.3	3.6	5.1
InstructBLIP (Dai et al., 2023)	5.7	9.6	5.7	5.7	6.9
GPT-4V (OpenAI, 2024)	6.0	9.9	5.4	8.4	3.9

4 Analysis

4.1 What is the Impact of Introducing the DSG?

To reasonably increase the difficulty of hallucination assessment, we introduced the DSG mechanism. As introduced in Section3.3, by reorganizing the problem into a tree structure, the judgment of each leaf node depends on the correctness of the root node's judgment. In this section, we quantitatively analyze the impact brought by the DSG.

Table 5 shows the changes in metrics for each model before and after introducing DSG. It can be seen that stronger models like GPT-4V and LLaVA are less affected, while the metrics for other models have dropped by more than half. The reason might be that these weaker models do not perform well on object-level questions. Therefore, after the introduction of the DSG, they are marked as failed on all related leaf node questions, leading to a significant impact.

4.2 Is the Information Extracted from Images More Comprehensive?

As shown in Figure 1, we extract information from both the image and the caption to construct QA pairs. Typically, the image itself contains more abundant information. In this section, we will verify whether the information extracted from the image is more comprehensive and diverse.

We have separately counted six indicators related to image and caption, mainly focusing on the three directions of object, attribute, and relation. As shown in Figure 4, it is evident that the information extracted from the image surpasses the other in both comprehensiveness and richness. This indicates that FIHA has successfully extracted more fine-grained information from the images, aligning with expectations and demonstrating the rationale and effectiveness of FIHA's methods.

4.3 Why are Our Benchmark Results Lower Than Others?

It's easy to see that our test results are lower than others, indicating that FIHA can detect more difficult and distinct issues. We analyzed that there are mainly three reasons: firstly, we added a large number of misleading negative samples, and since the model tends to give affirmative answers (Section 3.2.1), this increased the difficulty of evaluation. Secondly, the role of DSG directly impacts the results (2.3). Finally, the comprehensiveness of FIHA is more challenging than methods that focus

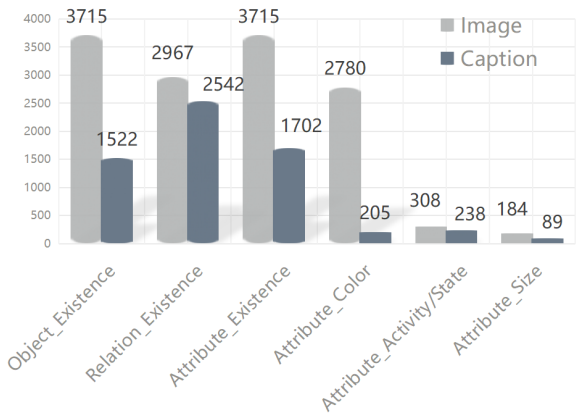


Figure 4: Comparison of the richness of information extracted from image and caption.

primarily on object-level.

Table 6: The accuracy of QA pairs generated from different datasets.

Dataset	From Image	From Caption
MS COCO (500)	98.2	96.0
No Foggy (50)	98.1	96.1
Medium Foggy (50)	97.6	94.5
Dense Foggy (50)	96.3	94.1

4.4 How Reliable is the Benchmark Generated by the Framework?

To test the how reliable is the benchmark, we check the accuracy of QA pairs with assist of YOLOv8 (Reis et al., 2024) and manully check. First, it can be observed from the results presented in Table 4 that the QA pairs generated from the captioning are highly reliable, achieving a 96% accuracy rate in sample from MSCOCO datasets , the QA pairs is 100% consistent with the captions. However, due do Blip2 can sometimes encounter hallucination, QA pairs sometimes generated QA which answers do not match the questions. Besides, questions and answers generated directly from images posed a few challenges. The pipeline use Bottom-up Attention (Anderson et al., 2018) and Fast R-CNN (Girshick, 2015) can perform with a 98.2% precision rate in detecting objects from everyday scenes—similar to the type of images in the MS COCO dataset—the remaining 1.8% error can still create incorrect QA pairs. For example, it sometimes fails to identify specific details, such as ears, in an image. Another challenge is feature extraction, for example, when the color of a horse was identified as black, but the horse was white or light gray. These rare inaccuracies represent one of the weak points of this framework, emphasizing the need for ongoing improvement in object detection and feature extraction technologies. Overall, the reliability of the FIHA in generating datasets to evaluate the hallucinations of LVLMs is remarkably high; the dataset generated from captions performs exceptionally well, with perfect accuracy.

4.5 How Robust and Generalizable is the Framework?

To evaluate the robustness and generalization ability of the FIHA framework's generated from image appproach, we conducted tests on complex scenes. Specifically, we used Foggy Cityscapes datasets (Cordts et al., 2016) in challenging conditions. We select 150 images in total with 50 images with three foggy level: dense, medium and no foggy to compare the influence of noise to the accuaracy of framework. For the QA generated from images approach: As shown in Table 6, in the dense condition, the accuracy of problem 96.3% for QA generated from images and 94.1% for QA generated from captions while medium is 97.6% and 94.5%, no foggy is 98.1% and 94.1%. The no foggy result remain same as mscoco. We observe that the model showed less confidence overall while it maintained over 99% confidence level for the main object but significantly lower confidence for minor or surrounding objects. Extracting feature such as color. This is becuase adverse weather degrades image quality, making accurate identification difficult. There are lots of previous study on object detection on sepcial scene (Wang et al., 2022). In foggy weather scenarios, the scattering and absorption of light by water droplets and particulate matter cause object features in images to become blurred or lost, presenting a significant challenge for target detection. Some feature such as color become invisble. For example, both the LLM and frcnn is easily to misrecongize the sliver into white. In summary, the FIHA framework is sufficiently robust and generalizes well, maintaining accurate classification with multiple data sets and in challenging image scenarios. It is able to adapt to almost any dataset to produce question-answer pairs for the assessment of hallucinations.

5 Related Work

In this section, we mainly discuss existing Large Vision Language Models (LVLMs) and the halluci-

nation problems that exist in LVLMs.

5.1 Large Vision Language Model

With the success of pretraining techniques in Large Language Models (LLMs) and Vision Foundation Models (VFMs), many researchers (Alayrac et al., 2022; Li et al., 2023a) have been expanding language models to comprehend real-world images through LVLMs with in-context or few-shot learning capabilities. As a result, there has been a surge in visual instruction-adapted LVLMs (Liu et al., 2023b; Zhu et al., 2023; Dai et al., 2023; Gong et al., 2023), demonstrating remarkable generalization performance across various Vision-Language (VL) tasks. Most of these studies utilized GPT-4 to generate multimodal instruction tuning datasets and multi-stage pretraining to align the visual information with the pretrained LLM. For example, Liu et al. (2023b) utilized the visual encoder output as input for LLaMA (Touvron et al., 2023) and trained both networks to align on the generated visual instruction dataset. Zhu et al. (2023) integrated Vicuna (Peng et al., 2023) as a language decoder and only fine-tuned the cross-modal alignment network with extended image captions from ChatGPT. Likewise, both Gong et al. (2023) and Dai et al. (2023) used various instruction-tailored VL datasets. However, the former adopted BLIP2 (Li et al., 2023b) as its foundational architecture while the latter initialized from Flamingo (Alayrac et al., 2022).

Despite the advancements of LVLMs, they remain encumbered by the persistent challenge of hallucinations when generating textual output. These issues significantly hinder their effectiveness in various vision-language tasks (Rohrbach et al., 2018).

5.2 Hallucination in LVLMs

Recently, there has been growing research attention directed towards the phenomenon of hallucination in LVLMs. Among these works, some studies, as shown in Table 1, have concentrated on hallucination detection and evaluation (Li et al., 2023d; Wang et al., 2023b,a; Jing et al., 2023), and some have developed methods to mitigate hallucination (Liu et al., 2023a; Zhou et al., 2023; Yin et al., 2023; Jing and Du, 2024). Though the issue of hallucination is studied extensively, only a few works have focused on fine-grained hallucination detection in LVLMs. For instance, Li et al. (2023d) proposed a novel evaluation metric "POPE" to evaluate hallucinations in LVLMs by pooling questions about the ground truth objects. They showed that existing state-of-the-art LVLMs are highly prone to object-level hallucinations. Wang et al. (2023b) introduced "HaELM," a framework for detecting hallucinations. They utilized LLM to generate a hallucinatory dataset and then fine-tuned LLaMA to identify hallucinatory responses from LVLMs. The aforementioned line of research either exclusively focused on object-level hallucination or required training for the detection of hallucination. To address these challenges, Wang et al. (2023a) introduced "AMBER," a comprehensive benchmark capable of assessing both generative and discriminative tasks, such as object attribute and relation hallucination. Though this work developed a fine-grained hallucination framework, it required human annotators to annotate the object existence, object attribute, and object relation information for discriminative tasks.

In contrast to the aforementioned studies, our work differs by being applicable to any existing dataset or unseen images for generating probing questions related to object existence, attributes, and relations for evaluating LVLMs hallucination. Instead of relying on human annotators, we use an object detection model that performs better in object detection tasks than LLMs. Our work does not require any additional information for an image to generate probing questions.

6 Conclusion

In recent years, large vision language models have developed quickly, but hallucinations remain a serious concern. Current hallucination evaluation methods face problems like high costs, limited scope, and lack of generalization. Thus, we introduce FIHA, a multi-dimensional detection method that requires no LLMs and no annotations. FIHA can automatically create high-quality QA pairs for any image dataset. We conducted a thorough analysis of the performance of mainstream LVLMs, identified the issues, and proposed potential methods for improvement. In the future, we will delve deeper into methods for alleviating hallucinations.

Liminations

FIHA has comprehensive features and maintains a high overall quality. Despite the limitations discussed in the previous analysis section, there are additional constraints in some aspects. The generated QA primarily focuses on the existence, attributes,

and relations of main objects in the images, while lacking in QA for surrounding and minor objects. This is due to the FRCNN's lower confidence in detecting small and less obvious objects.

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, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

Xiang Chen, Chenxi Wang, Yida Xue, Ningyu Zhang, Xiaoyan Yang, Qiang Li, Yue Shen, Jinjie Gu, and Huajun Chen. 2024. Unified hallucination detection for multimodal large language models. *arXiv preprint arXiv:2402.03190*.

Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. 2023. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-image generation. *arXiv preprint arXiv:2310.18235*.

Yuren Cong, Michael Ying Yang, and Bodo Rosenhahn. 2022. Reltr: Relation transformer for scene graph generation.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Ross Girshick. 2015. Fast r-cnn. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448.

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. Multimodal-gpt: A vision and language model for dialogue with humans. *CoRR*, abs/2305.04790.

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

Hongyu Hu, Jiyuan Zhang, Minyi Zhao, and Zhenbang Sun. 2023a. CIEM: Contrastive instruction evaluation method for better instruction tuning. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. 2023b. TIFA: accurate and interpretable text-to-image faithfulness evaluation with question answering. *CoRR*, abs/2303.11897.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.

Liqiang Jing and Xinya Du. 2024. Fgaif: Aligning large vision-language models with fine-grained ai feedback. *arXiv preprint arXiv:2404.05046*.

Liqiang Jing, Ruosen Li, Yunmo Chen, Mengzhao Jia, and Xinya Du. 2023. Faithscore: Evaluating hallucinations in large vision-language models. *arXiv preprint arXiv:2311.01477*.

Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. Longeval: Guidelines for human evaluation of faithfulness in long-form summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 1642–1661. Association for Computational Linguistics.

Ranjay Krishna, Michael S. Bernstein, and Li Fei-Fei. 2019. Information maximizing visual question generation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 2008–2018. Computer Vision Foundation / IEEE.

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

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. *arXiv preprint arXiv:2301.12597*.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023c. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023d. Evaluating object hallucination in large vision-language models. *CoRR*, abs/2305.10355.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023a. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 1(2):9.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning.

Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. 2023. Negative object presence evaluation (nope) to measure object hallucination in vision-language models. *arXiv preprint arXiv:2310.05338*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *CoRR*, abs/2305.14251.

OpenAI. 2024. Gpt-4 technical report.

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

Dillon Reis, Jordan Kupec, Jacqueline Hong, and Ahmad Daoudi. 2024. Real-time flying object detection with yolov8.

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

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2023. Aligning large multimodal models with factually augmented RLHF. *CoRR*, abs/2309.14525.

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*.

Hai Wang, Yansong Xu, Youguo He, Yingfeng Cai, Long Chen, Yicheng Li, Miguel Angel Sotelo, and Zhixiong Li. 2022. Yolov5-fog: A multiobjective visual detection algorithm for fog driving scenes based on improved yolov5. *IEEE Transactions on Instrumentation and Measurement*, 71:1–12.

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

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. 2023b. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. 2023. mplug-owl: Modularization empowers large language models with multimodality. *CoRR*, abs/2304.14178.

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

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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. *CoRR*, abs/2304.10592.

A Details of the Experimental Setup

Datasets We used two datasets: the MSCOCO(Lin et al., 2014) and the Foggy Cityscapes(Cordts et al., 2016). **MSCOCO** is a large image dataset developed by Microsoft, officially known as Microsoft Common Objects in Context. This dataset aims to advance the development of computer vision tasks such as object detection, segmentation, and image captioning. This dataset contains over 330,000 images, of which more than 200,000 images are annotated, covering 80 different object categories. **Foggy Cityscapes** is a synthetic fog dataset that simulates fog in real-world scenes. Each foggy image is rendered using clear images and depth maps from Cityscapes. Consequently, the annotations and data split in Foggy Cityscapes are inherited from Cityscapes.