FIHA: Automated Fine-grained Hallucinations Evaluations in Large Vision Language Models with Davidson Scene Graphs

Bowen Yan* **Zhengsong Zhang*** **Liqiang Jing*** **Eftekhar Hossain** **Xinya Du†**
University of Texas at Dallas, Richardson, United States
{bowen.yan, zhengsong.zhang, liqiang.jing, xinya.du}@utdallas.edu

Abstract

The rapid development of Large Vision-Language Models (LVLMs) often comes with widespread hallucination issues, making cost-effective and comprehensive assessments increasingly vital. Current approaches mainly rely on costly annotations and are not comprehensive – in terms of evaluating all aspects, such as relations, attributes, and dependencies between aspects. Therefore, we introduce the FIHA (autonomous Fine-graIned Hallucination evAluation in LVLMs), which could access LVLMs hallucination in an LLM-free and annotation-free way and model the dependency between different types of hallucinations. FIHA can generate Q&A pairs on any image dataset at minimal cost, enabling hallucination assessment from both image and caption. Based on this approach, we introduce a benchmark called FIHA-v1, which consists of diverse questions on various images from three datasets. Furthermore, we use the Davidson Scene Graph (DSG) to organize the structure among Q&A pairs, in which we can increase the reliability of the evaluation. We evaluate representative models using FIHA-v1, highlighting their limitations and challenges. We released our code and data at https://github.com/confidentzzzs/FIHA.

1 Introduction

Large Vision-Language Models (LVLMs) such as MiniGPT-4 (Zhu et al., 2023) and LLaVA (Liu et al., 2023c), which extend Large Language Models (LLMs) by incorporating visual encoders, such as CLIP (Radford et al., 2021), 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; Jing et al., 2024; Zhang et al., 2025). Specifically, the models may describe objects that do not exist in the image or incorrect object attributes and relations between objects. 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 (*i.e.,* Q&A pairs) for hallucination evaluation. 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, detailed explanations can be found in the Appendix C. (2) Additionally, most prior work heavily relies on human annotations (Wang et al., 2023a) or LLMs (Li et al., 2023c) to generate Q&A 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 assessing fine-grained and diverse hallucinations in large-scale vision-language models. The framework accepts either images or captions as input and generates Q&A pairs by extracting objects, attributes, and entity relations. It then formulates diverse questions (e.g., "what", "who", "which", etc.) that allow for free-form responses. By integrating BLIP-2 (Li et al., 2023b) for caption generation, Fast R-CNN (Girshick, 2015) for feature extraction, and a question-generation template, our pipeline enables fully automatic Q&A generation without relying on LLMs or manual annotations.

* Equal Contribution
† Corresponding Author

Findings of the Association for Computational Linguistics: ACL 2025, pages 12014–12026

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

Evaluation Methods	Discriminative Hallucination			Task Type		Dependency	LLM Free	Annotation Free
	Object	Attribute	Relation	Dis.	Gen.			
POPE (Li et al., 2023d)	✓	×	×	✓	×	×	✓	✓
NOPE (Lovenia et al., 2023)	✓	×	×	✓	×	×	×	✓
CIEM (Hu et al., 2023)	✓	✓	×	✓	×	×	×	✓
Bingo (Cui et al., 2023a)	×	×	×	×	✓	×	✓	×
AMBER (Wang et al., 2023a)	✓	✓	✓	✓	✓	×	✓	×
HallusionBench (Liu et al., 2023a)	×	×	×	×	✓	×	✓	×
MHaluBench (Chen et al., 2024)	✓	✓	×	×	✓	×	×	×
Hal-Eavl (Jiang et al., 2024a)	✓	✓	✓	✓	✓	×	×	✓
FIHA (ours)	✓	✓	✓	✓	✓	✓	✓	✓

We organize Q&A pairs using the Davidson Scene Graph (DSG) (Cho et al., 2023). The DSG ensures leaf node responses depend on root node answers, reducing errors and improving reliability. Our Q&A pairs include negative, narrative, and interrogative questions, allowing a progressive, comprehensive evaluation of image understanding.

We make the following key contributions through this work:

- To the best of our knowledge, FIHA is the first automated hallucination evaluation framework 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 DSG-based fine-grained evaluation benchmark FIHA-v1 that includes Q&A 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 Methodology

In this section, we introduce the overall pipeline of FIHA as illustrated in Figure 1. In summary, our pipeline offers two approaches for Q&A generation. The first is based on images: given an image I, we extract the necessary entities, including features such as objects, object attributes, and relations. Using a rule-based method, we then generate the Q&A pairs. The second approach is based on captions. If an image does not already have a caption, we can use BLIP-2 (Li et al., 2023b) to generate captions. Alternatively, if the dataset includes original captions, we can use them directly as input, pass them through the feature extraction process, and generate the corresponding Q&A pairs.

2.1 Fine-grained Information Extraction

2.1.1 Information Acquisition from the Caption

Q&A generation based on caption includes caption generation (optional if original datasets include captions) and extract information (object existence, object attributes and object relations) and using these information to generate Q&A pairs.

Caption Generation. Image captions can depict an image in detail, demonstrating fine-grained visual information, such as objects, attributes and relations. Inspired by the findings of previous research (Li et al., 2023d), which indicate that smaller vision-language models tend to produce more concise responses with fewer hallucinations compared to mainstream LVLMs. As such, we select BLIP-2 to generate a caption for the image. This step allows us to generate highly credible captions based on the image.

Fine-grained Information Extraction. In this process, we take either the generated captions (if the ground-truth caption is not available) or the ground-truth captions, depending on the user's needs, as input and extract information such as object existence, object attributes, and relations from the captions. For extracting objects and attributes in the caption, we use SpaCy's (Honnibal and Montani, 2017) part-of-speech tagging feature to identify objects and their corresponding attributes, such as numerals, adjectives, and verbs. As a result, we obtain all the ground truth objects and their attributes as: $G^C_{O,A} = \{o_1 : A_1, o_2 : A_2, \ldots, o_n : A_n\}$, where n is the number of objects. o_i is the i-th object and A_i is all attributes for the i-th object. Relations from the captions are extracted

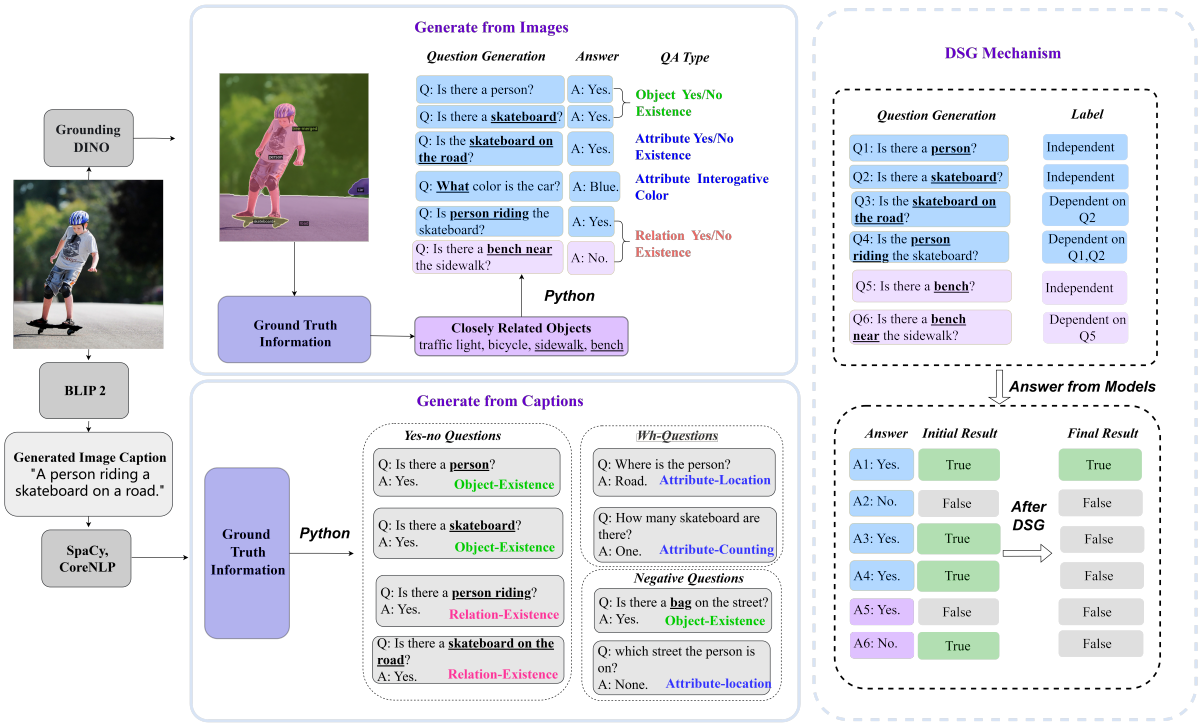


Figure 1: Overview of FIHA framework. FIHA extracts entities, attributes, and relations from images and captions respectively, and generates comprehensive and diverse questions to thoroughly detect hallucinations. In the Figure, we can see that no large language model (LLM) (Achiam et al., 2023) or additional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

using the Stanford CoreNLP library[1], which provides a powerful suite of NLP tools for performing various linguistic analyses on text, making it an ideal choice for relation extraction. From this process, we obtain all the relations: $G_R^C = \{R_1(o_{R_1}^1, o_{R_1}^2), \ldots, R_m((o_{R_m}^1, o_{R_m}^2))\}$, where m is the number of relations. R_i is the i-th relation for the objects $o_{R_i}^2$ and $o_{R_i}^2$.

2.1.2 Information Acquisition from the Image

As the caption may lose some information in the image, our second approach to extract information is directly from images, which provides richer and more detailed information than captions alone. For object and attribute extraction, we use Grounding DINO (Liu et al., 2023d), a well-established and widely used object detection method based on Transformer architecture. Grounding DINO has been a pioneering approach in the field of object detection due to its ability to quickly identify objects within an image while simultaneously predicting their attributes. This method allows us to retrieve the ground truth objects along with their corresponding attributes such as color, size, and shape, forming a set of objects and attributes: $G_{O,A}^I = \{o_1 : A_1, o_2 : A_2, \ldots, o_n : A_n\}$, where n represents the number of objects detected. In addition to identifying objects and their attributes, it is crucial to understand how these objects interact or relate within the scene. For this purpose, we employ RelTR (Cong et al., 2022), a cutting-edge method designed to generate sparse scene graphs by decoding visual appearances and learning both subject and object queries from the image data. RelTR enables us to extract meaningful relationships between the detected objects, such as spatial relations (e.g., one object being behind or near another) and actions (e.g., wearing or holding), resulting in a set of ground truth relations: $G_R^I = \{R_1(o_{R_1}^1, o_{R_1}^2), \ldots, R_m((o_{R_m}^1, o_{R_m}^2))\}$, where m denotes the number of relationships extracted from the image. By combining both object detection and relational extraction, this approach provides a comprehensive understanding of the visual content, which is essential for generating accurate and meaningful Q&A pairs.

2.2 Question Answer Pair Generation

Next, we generate corresponding Q&A pairs for the extracted image and text information, respectively.

[1] https://stanfordnlp.github.io/CoreNLP/.

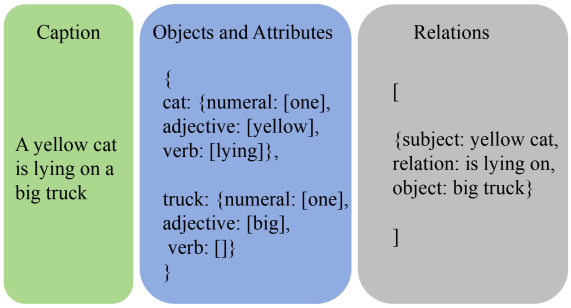


Figure 2: Example of extracted information.

After reorganizing them through DSG, they can be directly input into the model to detect the level of hallucination.

We use two kinds of questions for hallucination evaluation. The first type is Yes-No questions, which check object existence, such as "Is there any $\{\text{obj}_k\}$?" and object relations like "Is there a $\{\text{obj}_1\}$ near the $\{\text{obj}_k\}$?". These questions help determine whether specific objects and their relationships are present within an image. Additionally, Yes-No questions assess attributes by asking about features like color, size, or location.

The second type is Wh-Questions, which add diversity to the evaluation by incorporating interrogative words such as "what", "who", "which", "where", and "how many". These questions elicit more detailed, free-form responses, typically no longer than three words. For example, "What color is the $\{\text{obj}_k\}$?" or "Which object is near the $\{\text{obj}_k\}$?" help assess the finer details about objects and their relationships. Unlike traditional hallucination evaluations that primarily rely on Yes-No questions, our approach includes both types to provide a more comprehensive assessment.

We introduce Negative Questions for both Yes-No and Wh-Questions. These questions are created by replacing real objects, attributes (e.g., color, size), and relations in the original Q&A pairs with non-existent ones. For example, "Is there a car here?" becomes "Is there a plane here?" using a randomly selected object from a pre-constructed set. To ensure accurate evaluation, we avoid including objects present in the current image in the candidate set. These questions are answered with negative pronouns like "none," "nobody," or "nowhere."

2.3 Davidson Scene Graph

To model the dependency between objects, attributions, and relations accurately and improve the reliability of hallucination evaluation, we introduce the Davidsonian Scene Graph (DSG) (Cho et al., 2023) mechanism. The DSG schematic diagram can refer to Figures 1 and 8. The DSG can be understood as a post-processing step for the Q&A pairs. After obtaining all the Q&A pairs, we organize them into multiple tree-like structures, where each Q&A pair serves as a node. According to the structure of the tree-like structures, each node is either a root node or a leaf node. Specifically, the entire process is divided into three steps. In step 1, we set the question about the existence of a certain object as the root node. In step 2, we set all questions related to the object of the root node, such as those about its attributes and relations, as corresponding leaf nodes. Finally, in step 3, determine whether the root node question is answered correctly; if not, there is no need to judge the questions at the leaf nodes, and we directly determine that all questions on the tree are answered incorrectly. For instance, after step 1 and step 2, we obtain a list of questions such as $L^Q = \{Q_1 : Independent, Q_2 : Depends\ on\ Q_1\}$. Before determining if the answer to Q_2 is correct, we first assess Q_1, which concerns the accuracy related to the root node. If the question about the existence of an object, which is at the root node, is answered incorrectly, we consider that all other related questions must be hallucinatory.

Table 2: The number of questions generated by FIHA from various datasets.

Source	From Image	From Caption
MSCOCO	25,699	13,007
Foggy	7,232	2,801
Visual Genome	1,566	476

3 Experiments

3.1 Setup

Datasets. We construct a hallucination evaluation benchmark FIHA-v1 based on three datasets: the MSCOCO (Lin et al., 2014), the Foggy (Cordts et al., 2016) and Visual Genome (Krishna et al., 2017). **MSCOCO** is a large image dataset by Microsoft with over 330,000 images. More than 200,000 are annotated across 80 object categories. **Foggy** is a synthetic fog dataset with 1,500 images, each in three fog levels (no fog, medium fog, dense fog). **Visual Genome** has 108,077 images with some overlap with MSCOCO. For our benchmark, we only use the test sets of these datasets to avoid

Table 3: Evaluation results of LVLMs on questions generated from images and captions using FIHA. The upper part is from the MSCOCO dataset and the bottom part is from the Foggy dataset. F1 (Gen) is the BERTScore value calculated from the standard answers and model outputs for all Wh-Questions. For more details, please refer to the explanation of Metrics in Section 3.1.

Model	Question Generated from Image					Question Generated from Caption				
	Acc.	P.	R.	F1	F1 (Gen)	Acc.	P.	R.	F1	F1 (Gen)
MSCOCO										
mPLUG-Owl	42.1	70.2	61.4	43.7	15.2	31.4	61.6	55.5	31.2	11.4
MiniGPT-4	23.5	27.5	22.2	22.1	21.6	15.9	25.7	28.8	14.2	18.4
MultiModal-GPT	59.1	46.4	47.1	46.6	16.1	23.8	39.6	45.7	22.1	10.8
LLaVA-1.5-7B	77.8	77.0	65.9	67.7	21.4	50.7	64.9	67.5	50.5	13.7
LLaVA-1.5-13B	78.9	80.9	66.4	68.3	20.9	47.6	64.2	65.5	48.5	13.8
InstructBLIP	84.7	83.3	78.6	80.4	21.8	65.7	69.5	77.4	64.2	14.1
GPT-4V	87.2	81.4	86.3	85.5	25.2	70.3	71.5	75.8	69.3	22.7
Foggy										
mPLUG-Owl	64.8	60.2	51.1	42.7	18.6	29.5	58.9	51.6	25.6	29.3
MiniGPT-4	30.1	30.2	27.6	28.1	9.4	23.4	34.4	37.8	23.0	11.6
MultiModal-GPT	50.2	48.7	46.1	45.8	17.6	28.1	43.9	47.9	25.4	24.5
LLaVA-1.5-7B	67.7	68.4	56.2	52.9	19.7	29.1	50.0	49.2	25.8	27.5
LLaVA-1.5-13B	68.1	71.5	56.1	52.3	18.8	28.9	49.2	49.8	25.5	27.7
InstructBLIP	70.9	75.6	60.2	58.8	20.3	32.8	58.3	53.2	30.5	29.2
GPT-4V	76.3	70.1	64.6	66.0	16.2	33.7	53.3	51.7	32.1	21.7
Visual Genome										
mPLUG-Owl	41.8	68.9	60.9	43.3	16.4	44.6	71.7	51.8	33.8	24.1
MiniGPT-4	22.9	26.8	22.0	21.8	22.3	15.9	25.7	28.8	14.2	16.8
MultiModal-GPT	58.8	46.1	46.9	46.3	17.2	65.3	65.2	62.2	61.7	21.7
LLaVA-1.5-7B	77.7	77.2	61.1	67.9	20.6	56.4	73.8	62.0	52.6	20.1
LLaVA-1.5-13B	79.0	81.2	66.7	68.6	20.9	74.3	79.6	68.8	69.2	20.2
InstructBLIP	84.5	83.7	79.0	80.7	22.4	67.7	78.4	71.9	66.7	26.8
GPT-4V	87.0	81.2	86.0	85.3	23.7	84.2	78.9	84.1	82.2	26.0

overlap with training data used by LVLMs.

Metrics. We use Accuracy (Acc.), Precision (P.), Recall (R.), and F1 Score (F1) as evaluation metrics for Yes-No questions. For Wh-Questions, we use the F1 (Gen) from BERTScore (Zhang et al., 2020) for evaluation.

Models. We select seven mainstream LVLMs for evaluation: mPLUG-Owl (Ye et al., 2023), MiniGPT-4 (Zhu et al., 2023), MultiModal-GPT (Gong et al., 2023), LLaVA-1.5-7B (Liu et al., 2023c), LLaVA-1.5-13B (Liu et al., 2023c), InstructBLIP (Dai et al., 2023), and GPT-4V (OpenAI, 2024).

3.2 Data Processing and Analysis

We randomly selected 500 images from the MSCOCO dataset, 150 images from the Foggy, and 50 from the Visual Genome dataset. Using the process described in Section 2, we generate tens of thousands of Q&A pairs. The detailed data statistics of our FIHA-v1 benchmark can be found in Table 2.

Figure 3 illustrates the distribution of question types generated from images and captions. The proportion of questions related to objects, attributes, and relations is relatively balanced, reflecting the rationality of the method design. It is noteworthy

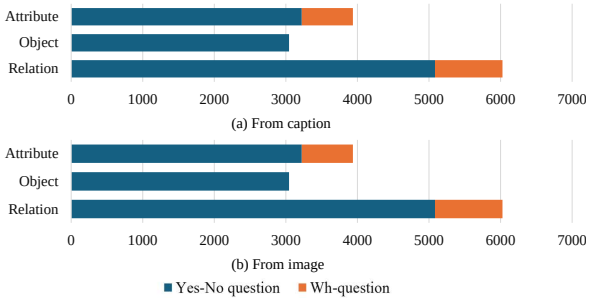


Figure 3: Distribution of two types of question, *i.e.,* Yes-No and Wh-questions.

that the abundance of the question category reflects FIHA's effective capability in generating tasks of the generation type, thereby enabling a more effective assessment of hallucinations.

3.3 Experimental Results

3.3.1 Overall Results on Datasets Generated by FIHA

We show the hallucination comparison of the seven mainstream LVLMs on our FIHA-v1 in Table 3. From this Table, we have several observations. 1) It's worth highlighting that GPT-4V excels in both image and caption Q&A pairs, achieving the best performance among the evaluated models. 2) The second-best performer is InstructBLIP, which sig-

Table 4: The fine-grained assessment of LVLMs evaluates object, attribute, and relation accuracy using Q&A pairs generated from captions in the MSCOCO and Foggy Cityscapes datasets.

Model	Object				Attribute				Relation			
	Acc.	P.	R.	F1	Acc.	P.	R.	F1	Acc.	P.	R.	F1
MSCOCO												
mPLUG-Owl	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	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	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	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	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	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	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												
mPLUG-Owl	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	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	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	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	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	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	61.8	69.6	59.2	54.5	11.1	37.0	33.1	11.0	20.4	50.5	50.4	20.3

nificantly outperforms other models except GPT-4V across most metrics. 3) Additionally, we have observed that model parameters are also significant factors affecting performance. For instance, LLaVA-1.5-13B provides a more comprehensive improvement over the LLaVA-1.5-7B.

In addition, we also show the performance of 7 mainstream LVLMs on FIHA-v1 based on the Visual Genome dataset. The results show a similar trend as compared to the performance in MSCOCO datasets. Specifically, the GPT-4V performs best and MiniGPT-4 performs the worst. LLaVA-1.5-13B performs better than LLaVA-1.5-7B, which also indicates that the model parameter size influences the performance.

3.3.2 Fine-Grained Results

Furthermore, we evaluate the model's performance from more dimensions (*i.e.,* the object existence, attribute, and relation) with FIHA. We show the fine-grained evaluation results in Table 4.

Object Hallucination From the results for the object, we can observe 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 From the results for the attribute, tt 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 From the results for the relation, this part is the most challenging, with the F1 score of GPT-4V on regular data not even reaching 60%. The potential reason is that Q&A pairs of the relation types involved more than one object, which makes it challenging.

4 Analysis

In this section, we further evaluate the effectiveness of our benchmark FIHA-v1 by four research questions.

4.1 How Reliable is FIHA?

To evaluate the benchmark's reliability, we manually check the accuracy of Q&A pairs in FIHA-v1 generated by the pipeline, verifying if the answers match the questions. In the human evaluation process, we employ annotators manually evaluate whether the answer is right in each image, question, and answer pair. Whether it is a Yes-No question or a Wh-Question, it will be marked as True or False, that is, whether the answer is right for the question. The final accuracy is calculated as (num_true / (num_true + num_false)). Table 5 shows that

Table 5: The results of human evaluation (accuracy) of Q&A pairs generated from different datasets.

Source	From Image	From Caption
MSCOCO	98.2	96.0
No Foggy	98.1	96.1
Medium Foggy	97.6	94.5
Dense Foggy	96.3	94.1

Q&A pairs from image captions are 96% accurate on MSCOCO samples. The MSCOCO-based pipeline using Grounding DINO achieves 98.2% accuracy, with 1.8% errors, such as missing details or misidentifying colors. Overall, FIHA shows high reliability in generating datasets for evaluating hallucinations in LVLMs, with near-perfect accuracy in caption-based datasets.

We also test on complex images using the Foggy dataset (Cordts et al., 2016), which test evaluates noise's effect on the framework's accuracy. Examples are in Appendix B.

As shown in Table 5, under dense fog, the accuracy for *Q&A pairs generation from images* is 96.3%, while for *Q&A pairs generation from captions*, it is 94.1%. For medium fog, the accuracies are 97.6% and 94.5%, respectively. Under no fog, the accuracies are 98.1% and 94.1%, similar to MSCOCO results.

The results show that as fog increases, the accuracy of FIHA's Q&A pairs decreases, highlighting the challenge of blurry images.

4.2 What is the Impact of Introducing DSG?

To improve hallucination assessment, we propose the DSG mechanism, which models dependencies between hallucination types. Table 6 shows that stronger models like GPT-4V and LLaVA-1.5-13B exhibit smaller performance drops (6.0% and 2.7%), indicating their robustness to dependency-based evaluations. In contrast, weaker models such as MultiModal-GPT and mPLUG-Owl show substantial declines (21.3% and 29.6%), reflecting frequent root-level errors that propagate to leaf-level questions, which these models might otherwise answer correctly under standard evaluation.

The results reveal that most hallucination errors occur at fundamental levels, such as object recognition, with models like LLaVA-1.5-13B maintaining high accuracy due to fewer cascading errors. Precision drops more than recall across models, suggesting that DSG effectively uncovers false positives. For example, MiniGPT-4's precision decreased by 11.0%, highlighting previously unnoticed errors.

Table 6: The performance decrease of various LVLMs after introducing DSG.

Model	Acc.↓	P.↓	R.↓	F1↓	F1 (Gen)↓
mPLUG-Owl	29.6	22.1	14.0	28.7	14.2
MiniGPT-4	62.6	51.8	62.1	61.2	42.3
MultiModal-GPT	21.3	27.6	21.9	24.3	12.9
LLaVA-1.5-7B	4.2	11.7	4.5	4.8	5.7
LLaVA-1.5-13B	2.7	8.1	3.3	3.6	5.1
InstructBLIP	5.7	9.6	5.7	5.7	6.9
GPT-4V	6.0	9.9	5.4	8.4	3.9

While InstructBLIP shows a significant F1 drop (from 9.6% to 5.7%), GPT-4V remains relatively stable (9.9% to 8.4%), suggesting stronger contextual reasoning. These results demonstrate that DSG provides a rigorous evaluation, exposing model weaknesses that standard assessments may miss.

4.3 How and Why LLM is Free?

FIHA has a big benefit: it doesn't need extra big language models like GPT-4. This is shown when we make questions from true information. We do this with Python code, and you can find more details in Section 2.2 and Appendix A.

We don't use big language models to help make questions because they cost a lot. For example, if someone wants to make questions for 500 pictures, they would need about 36,900 questions (like we saw with MS COCO). This would cost almost $400 in API fees, which is very expensive.

4.4 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 Q&A 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 than that extracted from the caption. We have separately counted the number of six different types of Q&A pairs from 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 information extracted from the caption.

4.5 Why are Performance on our Benchmark Lower Than Others?

In the experiment, we observe that our test results are lower than others, *e.g.,* POPE (Li et al., 2023d) and HaELM (Wang et al., 2023b), indicating that FIHA can detect more difficult and distinct issues. We analyze that there are mainly three reasons:

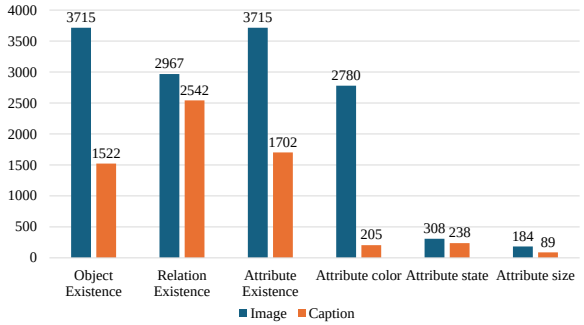


Figure 4: Comparison of the number of Q&A paris across different types of hallucination from image and caption.

firstly, we added a large number of misleading negative samples, and since the model tends to give affirmative answers (Section 3.3.1), this increases the difficulty of evaluation. Secondly, the role of DSG directly impacts the results and improves the reliability of the evaluation method. Finally, the comprehensiveness of FIHA is more challenging than methods that focus primarily on generating coarse-grained object-level questions.

4.6 Will Using Fixed Templates Limit the Diversity and Types of Questions?

According to the description of LVLM hallucinations in the existing work (Bai et al., 2024), the types of questions usually limited, even if the questions are generated by LLMs. We compared the method (Chen et al., 2024; Jiang et al., 2024b) of using LLMs to generate questions and found that the diversity of questions generated by the LLM-based method is similar to our method.

4.7 Why Are the Metrics Lower on the Foggy Dataset Compared to MSCOCO?

There are mainly two reasons: 1. The Foggy dataset is a collection of images captured in foggy weather conditions, which inherently falls under the category of complex scenes (such as nighttime, underwater, rainy conditions, etc.). Due to reduced visibility in foggy environments, many objects in the images become blurred, increasing the difficulty for models to accurately recognize them. 2. The training set of MSCOCO is commonly used for training various LVLMs, while datasets of the Foggy type are rarely used for training. Therefore, we can see Foggy as a out-of-distribution test setting. This results in the various LVLMs being unfamiliar with the Foggy style.

5 Related Work

In this section, we mainly discuss existing Large Vision-Language Models (LVLMs) and the hallucination problems that exist in LVLMs.

Large Vision-Language Model With the success of pretraining in LLMs (Touvron et al., 2023) and VFMs (Awais et al., 2023), many researchers (Alayrac et al., 2022; Li et al., 2023a) extended LLMs to understand real-world images through LVLMs, benefiting from in-context and few-shot learning. This led to a rise in visual instruction-adapted LVLMs (Liu et al., 2023c; Zhu et al., 2023; Dai et al., 2023; Gong et al., 2023), which show strong generalization across VL tasks. Most use GPT-4 to generate multimodal instruction datasets and multi-stage pretraining to align visual data with LLMs. For example, Liu *et al.* (Liu et al., 2023c) aligned LLaMA (Touvron et al., 2023) with a visual encoder output, while Zhu *et al.* (Zhu et al., 2023) fine-tuned Vicuna (Peng et al., 2023) for cross-modal alignment. Similarly, Multimodal GPT (Gong et al., 2023) and InstructBLIP (Dai et al., 2023) used VL datasets, with the former using BLIP2 (Li et al., 2023b) and the latter starting from Flamingo (Alayrac et al., 2022).

Despite these advances, LVLMs still struggle with hallucinations in textual output, limiting their effectiveness in vision-language tasks (Rohrbach et al., 2018).

Hallucination in LVLMs Recent studies have focused on hallucination in LVLMs. Some works, summarized in Table 1, address hallucination detection and evaluation (Li et al., 2023d; Wang et al., 2023b,a; Jing et al., 2023), while others propose mitigation methods (Liu et al., 2023b; Zhou et al., 2023; Yin et al., 2023; Jing and Du, 2024; Jing et al., 2025). For instance, Bingo (Cui et al., 2023b) evaluates GPT-4V's hallucinations with bias and interference, and HallusionBench (Guan et al., 2024) diagnoses entangled language hallucinations and visual illusions. AutoHallusion (Wu et al., 2024) generates benchmarks by manipulating images to challenge language priors, and Hal-Eval (Jiang et al., 2024b) categorizes hallucinations into objects, attributes, relations, and events.

Despite progress, fine-grained detection is less explored. Li *et al.* (Li et al., 2023d) introduced POPE to evaluate object-level hallucinations, showing LVLMs' susceptibility. Wang *et al.* (Wang et al., 2023b) proposed HaELM, creating a hallu-

cination dataset and fine-tuning LLaMA for detection. These methods focus on object-level issues or require training. To address limitations, Wang *et al.* (Wang et al., 2023a) developed AMBER, a benchmark for generative and discriminative tasks involving object, attribute, and relation hallucinations, though it depends on human annotations.

Existing methods mostly primarily use LLMs to extract key information from image captions, such as objects, colors, positions, etc. Based on the extracted keywords and different prompts, the LLM then generates corresponding questions. For each generated question, the LLM also provides the corresponding answer. It can be seen that every step relies entirely on the LLM and manually designed prompts. In contrast, our method does not rely on LLMs for any step, from key information extraction to question and answer generation. When generating questions, it fills in pre-defined, well-structured templates, ensuring both accuracy and controllable question types. Therefore, our method is effeicent and not costly.

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 Q&A 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 Q&A primarily focuses on the existence, attributes, and relations of main objects in the images, while lacking in Q&A 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.

Muhammad Awais, Muzammal Naseer, Salman H. Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. 2023. Foundational models defining a new era in vision: A survey and outlook. *CoRR*, abs/2307.13721.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

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)*.

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

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

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.

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.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14375–14385.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

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

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.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2024a. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27036–27046.

Chaoya Jiang, Wei Ye, Mengfan Dong, Hongrui Jia, Haiyang Xu, Ming Yan, Ji Zhang, and Shikun Zhang. 2024b. Hal-eval: A universal and fine-grained hallucination evaluation framework for large vision language models.

Liqiang Jing, Guiming Hardy Chen, Ehsan Aghazadeh, Xin Eric Wang, and Xinya Du. 2025. A comprehensive analysis for visual object hallucination in large vision-language models.

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*.

Liqiang Jing, Jingxuan Zuo, and Yue Zhang. 2024. Fine-grained and explainable factuality evaluation for multimodal summarization. *CoRR*, abs/2402.11414.

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

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

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023b. 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. 2023c. Visual instruction tuning.

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

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*.

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*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry andf Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

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

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

Junyang Wang, 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*.

Xiyang Wu, Tianrui Guan, Dianqi Li, Shuaiyi Huang, Xiaoyu Liu, Xijun Wang, Ruiqi Xian, Abhinav Shrivastava, Furong Huang, Jordan Lee Boyd-Graber, Tianyi Zhou, and Dinesh Manocha. 2024. Autohallusion: Automatic generation of hallucination benchmarks for vision-language models.

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*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yue Zhang, Liqiang Jing, and Vibhav Gogate. 2025. Defeasible visual entailment: Benchmark, evaluator, and reward-driven optimization. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 25976–25984. AAAI Press.

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 Code Example for Generating QA Pairs Based on Extracted Information

```
if relation.endswith(tuple(['ing', 'ed'])):
    question = f"Is the {subject} {relation} the {object} in the image?"
elif relation.endswith(tuple(['over', 'under', 'above', 'near', 'behind', 'on', 'at'])):
    if obj_is_living(object):
        question = f"Who is {relation} the {object} in the image?"
    else:
        question = f"What is {relation} the {subject} in the image?"
```

B Example of foggy Cityscapes Images datasets

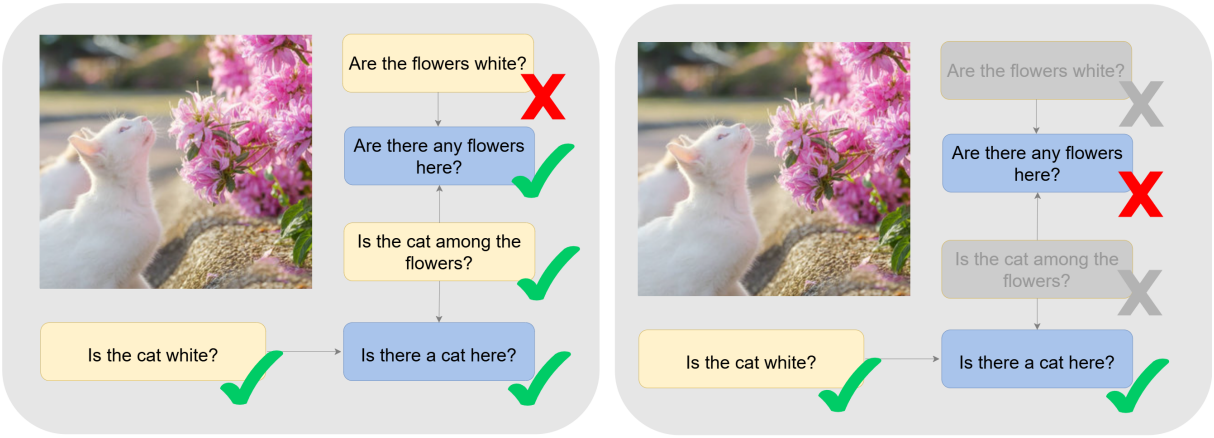
Figure 5: no foggy

Figure 6: medium foggy

Figure 7: dense foggy

C Expalnation for DSG

Depending on the answers, some questions in the hallucination benchmark become invalid and thus should not be asked to the LVLM to evaluate hallucination. As shown in Figure 8, if the answer to "are there any flowers here?" is no, dependent questions like "are the flowers white?" should be skipped – the LVLM may often say "the flower doesn't exist, but it is white".



(a) Score: 4/5 = 0.80

(b) Score: 2/5 = 0.40

Figure 8: The diagram on the right is a schematic illustration of the impact on the results after the introduction of DSG.