Unified Triplet-Level Hallucination Evaluation for Large Vision-Language Models

Anonymous authors
Paper under double-blind review

Abstract

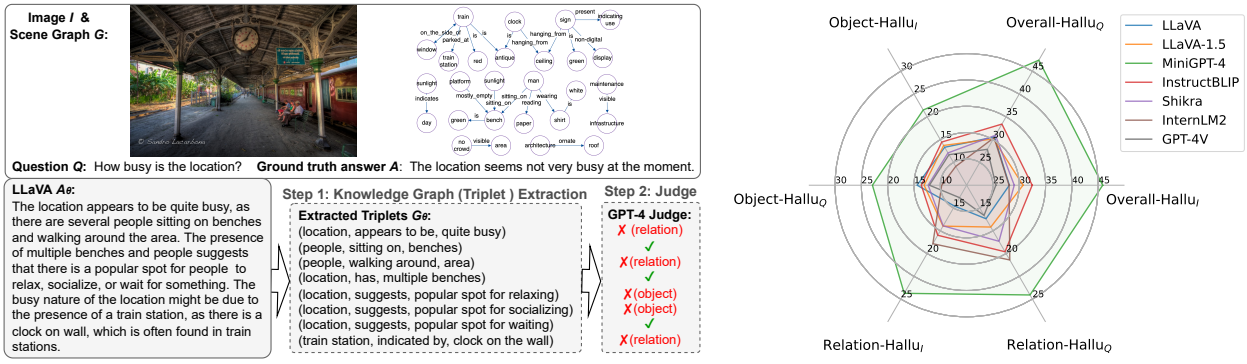
Despite the outstanding performance in vision-language reasoning, Large Vision-Language Models (LVLMs) might generate hallucinated contents that do not exist in the given image. Most existing LVLM hallucination benchmarks are constrained to evaluate the *object-related hallucinations*. However, the potential hallucination on the relations between two objects, *i.e.*, *relation hallucination*, still lacks investigation. To remedy that, we design a unified framework to measure object and relation hallucination in LVLMs simultaneously. The core idea of our framework is to evaluate hallucinations in (object, relation, object) triplets extracted from LVLMs' responses, making it easily generalizable to various vision-language tasks. Based on our framework, we further introduce **Tri-HE**, a novel **Tri**plet-level **H**allucination **E**valuation benchmark which can be used to study both object and relation hallucination at the same time. With comprehensive evaluations on Tri-HE, we observe that the relation hallucination issue is even more serious than object hallucination among existing LVLMs, highlighting a previously neglected problem towards reliable LVLMs. Moreover, based on our findings, we design a simple training-free approach that effectively mitigates hallucinations for LVLMs.

1 Introduction

Large Vision-Language Models (LVLMs) (Dai et al., 2023; Liu et al., 2023b; Chen et al., 2024a; Cai et al., 2024) have attracted significant attention. Despite the superior performances, existing works primarily focus on enhancing the *helpfulness* of LVLMs without careful consideration of the *reliability* of responses generated by LVLMs. However, it has already been observed by recent literature that LVLMs suffer from severe hallucination (Li et al., 2023d; Wang et al., 2023b;c; Guan et al., 2024; Chen et al., 2024b), *i.e., LVLMs might generate contents that do not exist in the given image*, probably due to insufficient training during visual instruction tuning. A typical example is provided in Figure 1a, where the LLaVA (Liu et al., 2023b) model considers the location to be busy, simply because LLaVA recognizes that it is a train station with several people existing.

With the prevalence of LVLMs, enormous works have started to explore the evaluation and analysis of LVLM hallucination. However, two problems are observed: 1) **Hallucination category:** most existing works focus on *object-related hallucination* (Li et al., 2023d; Wang et al., 2023b; Chen et al., 2024c) (*i.e.*, LVLMs describing an object not existing in the given image) while ignoring the possibility that even when two objects are successfully recognized, LVLMs might still mess up with their relationships when conducting commonsense reasoning. As illustrated in the example in Figure 1a, LLaVA successfully recognizes the "*people*" and the train station "*area*", yet predicts their relation to be "*walking around*" that cannot be directly obtained from the given image. Therefore, a unified definition and taxonomy is necessary to integrate different kinds of LVLM hallucination.

2) **Hallucination discrimination:** To evaluate how severe LVLMs hallucinate objects and their relationships within given images, prior works generally use either self-discrimination methods (*e.g.*, Yes/No questions) (Li et al., 2023d; Wang et al., 2023b; Guan et al., 2024; Wu et al., 2024) or template-driven discrimination approaches (*e.g.*, "*What is the relation with A and B?*") such as Reefknot (Zheng et al., 2024). However, such methods inherently constrain LVLMs to generate short answers like "*Yes/No*" or "*A has {} relation to*



(a) **Triplet-level LVLM hallucination evaluation pipeline.** (b) **LVLM hallucination comparison.**

Figure 1: **Overview of the unified hallucination evaluation pipeline**. (a) With the provision of images, scene graphs, and questions, knowledge graphs (*i.e.*, triplets) are extracted from LVLM responses, which are then judged by an LLM (GPT-4 here). (b) The radar plot showcases the evaluation results among different LVLMs (lower values demonstrate fewer hallucinations).

B". Given that LVLMs have varying capabilities to produce brief responses due to differences in pre-training datasets, this could introduce biases into the evaluation results. For instance, Chen et al. (2023a); Li et al. (2023d) have shown that InstructBLIP (Dai et al., 2023) tends to produce shorter outputs compared to other LVLMs, thus inflating its performance in answering the above type of questions and leads to hallucination evaluation bias. Moreover, these benchmarks require transforming general vision-language tasks into specific formats like "*Yes/No*", limiting their applicability. Therefore, we raise the following research question: **Can we develop a unified and unbiased evaluation framework capable of evaluating various types of hallucinations in LVLM responses across diverse tasks?**

To this end, we first propose a unified framework to simultaneously measure object and relation hallucinations in LVLM responses (§3). Specifically, our framework extracts knowledge graphs represented as triplets from LVLM-generated responses and then employs external evaluators to compare these triplets against the corresponding scene graphs from the input images. Consequently, our method facilitates hallucination evaluation for responses across diverse vision-language tasks, independent of specific question formats. Leveraging this unified framework, we further introduce **Tri-HE**, a novel benchmark for **Tri**plet-level **H**allucination **E**valuation, designed explicitly to assess both object and relation hallucinations (§4). Our experimental findings presented in §5 and Figure 1b confirm that relation hallucination poses a significant challenge for both closed-source and open-source LVLMs, often surpassing object hallucination in severity. By systematically comparing LVLMs' performance, we identify key insights that could potentially reduce hallucination rates (§5.2). Moreover, our proposed triplet-level hallucination judge, powered by LLMs, demonstrates impressive alignment with human judgments (Table 3). Motivated by these observations, we incorporate explicit triplet descriptions into LVLM prompts and introduce a straightforward yet effective training-free method to mitigate hallucinations (§5.4).

Our primary contributions are summarized as follows:

1. We propose a unified framework capable of jointly evaluating object and relation hallucinations in LVLM responses across diverse vision-language tasks. In particular, our triplet-level evaluation offers a finer-grained, more accurate assessment compared to existing methods.

2. Building upon this framework, we introduce **Tri-HE**, a novel triplet-level fine-grained hallucination evaluation benchmark tailored specifically for LVLMs.

3. We propose a simple yet highly effective training-free hallucination mitigation approach that surpasses the open-source LVLM competitors.

2 Related Work

2.1 Large Vision-Language Models (LVLMs)

The powerful capability exhibited by Large Language Models (LLMs) has facilitated the extension of LLMs towards the multi-modal domain. LLMs are empowered to understand and reason about both images and text by aligning representations from visual encoders to pre-trained language models, followed by visual instruction tuning. LLaVA (Liu et al., 2023a;b) proposes to use a simple projection layer to integrate the visual representations into textual encoders, which is further enhanced in Shikra (Chen et al., 2023b) by incorporating referential dialogue tasks. Instead, BLIP (Li et al., 2023a) proposes the Q-Former architecture to extract useful information from the visual representations, which is also used by MiniGPT-4 (Zhu et al., 2023) and InstructBLIP (Dai et al., 2023). InternLM (Dong et al., 2024) aligns with more diverse instruction data with the conditional online reinforcement learning from human feedback (RLHF) strategy, while MoCLE (Gou et al., 2023) further introduces the Mixture-of-Experts architecture into LVLMs to deal with the data conflict during instruction tuning. Although powerful, existing works primarily focus on improving the helpfulness, without a thorough analysis of the reliability of LVLMs.

2.2 Hallucination Evaluation in LVLMs

With the prevalence of LVLMs, a growing number of studies have been conducted on their hallucination issues (Chen et al., 2024b;d; Han et al., 2024; Huang et al., 2023; Li et al., 2023b; Wang et al., 2023b; Guan et al., 2024; Yue et al., 2024). Previous hallucination evaluation works can be categorized into two groups: 1) solely evaluating object hallucinations or do not distinguish different hallucinations (Zhao et al., 2023; Li et al., 2023c; Wang et al., 2023b; Chen et al., 2024c), which neglects other hallucination types like relation hallucination and is thus not comprehensive. The other type of works use "yes/no" questions to evaluate LVLM's relation/object hallucinations (Li et al., 2023d; Wang et al., 2023a; Guan et al., 2024; Wu et al., 2024). However, these benchmarks require transforming general vision-language tasks into "yes/no" formats, limiting their applicability. Also, different LVLMs may have different ability in answering such"yes/no" questions since they are pre-trained on different data, which may bias the evaluation results. To remedy this research gap, our paper proposes a triplet-level evaluation framework that can provide fine-grained object and relation hallucinations for responses to any vision-language tasks, with an evaluation benchmark Tri-HE that incorporates questions requiring more complicated commonsense reasoning.

It is noteworthy that a concurrent benchmark, Reefknot (Zheng et al., 2024), similarly assesses relation hallucinations at the triplet level. However, Reefknot exhibits several limitations compared to Tri-HE. First, Reefknot constructs VQA questions based on a simple template, "*What is the relation between A and B?*", restricting both the variety of vision-language tasks that can be evaluated and the length of LVLM-generated responses, potentially introducing evaluation biases. In contrast, our framework is flexible enough to be applied to various vision-language tasks. Moreover, since the questions in Tri-HE are generated by GPT-4V, it can cover a wider range of relation types compared to template-based questions, thus providing more comprehensive evaluation results. Second, Reefknot relies solely on a single entailment-based hallucination discriminator, whereas Tri-HE leverages powerful LLM-based discriminators capable of accurately and simultaneously identifying both object and relation hallucinations, leading to more detailed hallucination evaluation results.

3 Unified Hallucination Evaluation Framework Formulation

Inspired by the relation extraction (Xiaoyan et al., 2023) tasks in NLP, in this section, we propose a unified framework to evaluate both object and relation hallucinations via the object-relation triplets (*i.e.*, $(\text{Object}_1, \text{Relation}, \text{Object}_2)$). Here the objects and relations can either be a word or a phrase with attributes. We start by defining object and relation hallucinations via triplets in §3.1, based on which, we define our evaluation metrics and pipeline in §3.2 and §3.3 separately.

3.1 Definitions

As illustrated in Figure 1a, we formulate our framework with the standard VQA setting (though they can be generalized to evaluate hallucinations in any vision-language tasks given available scene graph annotations, as discussed in §3.4). Specifically, considering an input image I, a corresponding question Q associated with image I, its ground truth answer A, and the answer A_θ predicted by an LVLM $A_\theta(\cdot|Q, I)$ parameterized by θ, we first define:

- $G = (V, E)$ as the *scene graph* of I, where V and E refer to all the objects existing in I and all the possible relations among existing objects, respectively.
- $G' = (V' \subseteq V, E' \subseteq E)$ as the *knowledge graph* that includes all the required objects and relations to answer Q.
- $G_\theta = (V_\theta, E_\theta)$ as the *knowledge graph* extracted from A_θ, where V_θ and E_θ include all the objects and all the possible relations among objects mentioned in A_θ.

Note that here all graphs can be converted to a set of triplets (*i.e.*, $G = \{(v_1, e, v_2)\}$, where $v_1, v_2 \in V$ and $e \in E$). A common nightmare in previous LVLM hallucination literature lies in the ambiguous discrimination between prediction **hallucinations** and **errors** (Ji et al., 2023). To obtain unbiased hallucination evaluation results, we separate them depending on **whether or not the wrongly generated objects or relations exist in the given image** I. Specifically, given a triplet $(v_1, e, v_2) \in G_\theta$, we have the following definitions,

- **Object hallucination**: if $v_1 \notin V$ or $v_2 \notin V$, suggesting A_θ includes an object not within I. For example, the triplet (*location*, *suggests*, *popular spot for socializing*) in Figure 1a encounter an object hallucination since the object " *popular spot for socializing*" cannot be obtained from V.
- **Relation hallucination:** if $v_1, v_2 \in V$ yet $e \notin E$, suggesting that A_θ correctly recognizes two related objects from I but pair them with a non-existing relation. For example, the triplet (*people*, *walking around*, *area*) in Figure 1a has a relation hallucination since the relation "*walking around*" cannot be obtained from G, despite that the objects are all in V.
- **Prediction error:** if $v_1, v_2 \in V$ and $e \in E$ yet $(v_1, e, v_2) \notin G$, suggesting A_θ correctly recognizes objects and relations from I, yet pairs in a wrong way.

3.2 Evaluation Metrics

With the above definition in hand, given the knowledge graph G_θ extracted from a model response A_θ, we calculate the hallucination rates of A_θ as the **proportion of hallucinated triplets** in G_θ. Most previous works (*e.g.*, POPE (Li et al., 2023d)) directly evaluate the hallucination rate at the object-level with respect to the total number of predicted objects, yet make their results **not comparable among LVLMs**, since different LVLMs might refer to different numbers of objects in their responses. To address this issue, we instead opt to calculate the hallucination rate in the question- and image-level. Specifically, we calculate two types of hallucination rates, including the *question-level hallucination rate* (Hallu_Q) and *image-level hallucination rate* (Hallu_I), as defined in the following,

$$\text{Hallu}_\text{Q}(\{Q\}) = \frac{1}{|\{Q\}|} \left(\sum_{Q' \in \{Q\}} \left(\frac{\#\text{ HT in } G_\theta}{\#\text{ TT in } G_\theta} \right) \right) \times 100\%, \tag{1}$$

$$\text{Hallu}_\text{I}(\{I\}) = \frac{1}{|\{I\}|} \left(\sum_{I' \in \{I\}} \text{Hallu}_\text{Q}(\{Q_{I'}\}) \right) \times 100\%, \tag{2}$$

where HT is Hallucinated Triplets, TT is Total Triplets, $\{Q\}$ and $\{I\}$ are the sets of questions and images that LVLMs are evaluated on, respectively, and $\{Q_{I'}\} \subseteq \{Q\}$ suggest the subsets of questions related to the image I'. For both metrics, lower values demonstrate fewer hallucinations. Since the total number of questions and images is maintained the same for all evaluated LVLMs, $\textbf{Hallu}_\textbf{Q}(\cdot)$ **and** $\textbf{Hallu}_\textbf{I}(\cdot)$ **are indeed comparable and unbiased**.

3.3 Evaluation Pipeline

With the definitions and evaluation metrics provided in §3.1 and §3.2, the remaining problems contain two parts: 1) how to extract the knowledge graph G_θ from LVLM responses A_θ, and 2) how to judge a triplet in G_θ is hallucinated or not. The overview of our pipeline is illustrated in Figure 1a.

Knowledge Graph Extraction. Given an LVLM response A_θ with the corresponding question Q and image I, we extract the knowledge graph G_θ from A_θ via prompting GPT-4. Check our prompt for knowledge graph extraction in Appendix §A.1. Afterwards, we propose two different strategies to judge whether a triplet $(v_1, e, v_2) \in G_\theta$ includes hallucination based on the ground truth answer A and the image scene graph G, as described in the following.

NLI Judge. The first strategy is implemented with a natural language inference (NLI) (Reimers & Gurevych, 2019) model [1]. Specifically, given an extracted triplet, we first calculate its cosine similarity scores with all triplets in the image scene graph G and only retain those ground truth (GT) triplets with similarity scores greater than 0.5 to refine the information that will be used for the NLI model. If no triplets in G meet this criterion, only the top three GT triplets with the highest similarity scores will be kept, which are then taken as ground truth inputs for the NLI model to make predictions. If the NLI score between the extracted triplet and ground truth triplets is lower than 0.6 [2], suggesting the extracted triplet cannot be induced based on GT triplets, and therefore, resulting in a hallucination.

LLM Judge. Another evaluation strategy is to leverage prompting of a powerful LLM, a widely-adopted practice in recent works for assessing LLM outputs (Zheng et al., 2023). In this work, we primarily utilize GPT-4 in LLM judge to determine whether a given extracted triplet $(v_1, e, v_2) \in G_\theta$ can be **directly obtained** or **inferred** from the image scene graph G. Note that:

1. We do not employ GPT-4V in LLM judge, as Li et al. (2024) have reported that the text-only GPT-4 is more consistent with human preferences than GPT-4V.
2. Open-source models, such as LLaMA-3.3, can similarly deliver reliable and cost-efficient hallucination evaluation results (see detailed analysis in Table 3).

Additionally, if a triplet (v_1, e, v_2) is judged as hallucinated, we further prompt the LLM to clarify whether the hallucination pertains specifically to the relation e or the objects v_1, v_2. Refer to Appendix §A.2 for the prompt of LLM judge in our experiments [3].

3.4 Generalizability of our Framework

Although in the aforementioned sections, we formulate our unified hallucination evaluation framework primarily based on VQA tasks, it is capable of evaluating hallucinations in LVLM responses for any vision-language task, provided that corresponding scene graphs for the test images are available. This underscores the task-agnostic design of our proposed framework and highlights its strong generalization capability.

4 Tri-HE Construction

Following the formulation in §3, in this section, we provide a detailed discussion on how to construct our benchmark Tri-HE for a unified triplet-level evaluation of both hallucinations in LVLMs.

Image Collection. The construction of Tri-HE begins with images from the GQA dataset (Hudson & Manning, 2019), as the scene graph annotations provided by GQA naturally fit our triplet-level hallucination

[1] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[2] Check Appendix §B for how these thresholds are determined.

[3] For both knowledge graph extraction and LLM judge, we utilize the "gpt-4-1106-preview" model via OpenAI's API with default inference parameters.

#Images	#Questions	#Objects	#Relations	#Questions/Image	#Triplets/SG
300	1226	1723	618	4.09	19.10

Table 1: **Statistics of Tri-HE.** "SG" refers to Scene Graph.

evaluation formulation. Nevertheless, some scene graphs in GQA contain incomplete object relationships, omitting information necessary for accurate question answering. To mitigate this issue, we initially filter the GQA images, retaining only those whose scene graphs contain at least five object relations (edges between nodes). Subsequently, an annotator selects 300 images from the filtered images according to the following criteria:

1. Each image must contain more than two related objects.
2. Each image must be sufficiently clear to discern all visual details.

This procedure ensures a set of high-quality images suitable for subsequent dataset construction.

VQA Question Generation Next, since the VQA questions in the GQA dataset have already been extensively used during the pre-training of current LVLMs, we instead employ GPT-4V [4] to generate novel question-answer pairs for each image to avoid data contamination. To effectively examine both object and relation hallucinations in LVLM responses, we aim to generate questions that necessitate commonsense reasoning grounded on the provided images. Specifically, for every image, GPT-4V is prompted to generate 10 questions along with their answers [5], each requiring image-based commonsense reasoning to be answered. Furthermore, we ask GPT-4V to produce relation triplets describing the reasoning processes of answering the questions. These additional triplets can subsequently be used to enrich the original image scene graphs.

VQA Question Verification Following the initial generation of VQA questions, three annotators manually examine all generated questions, answers, and triplets based on the following criteria:

1. Each question must be valid and answerable using commonsense reasoning based on the provided image.
2. Each answer must appropriately address the question using commonsense reasoning.
3. Each triplet must accurately describe the corresponding answer and must only contain objects visible within the image.

Questions or answers failing to meet these conditions are discarded, while invalid triplets are excluded from the respective scene graphs. To validate annotation consistency, the annotators jointly annotate 100 question-answer pairs, achieving a Krippendorff's alpha (Krippendorff, 2011) of 0.62, demonstrating substantial inter-annotator agreement.

Statistics. The overall statistics for Tri-HE are summarized in Table 1. As described in Figure 2, each image in Tri-HE is linked to a scene graph and a set of question-answer pairs that require reasoning, accompanied by ground truth triplet annotations. Note that since the quality of each question in Tri-HE is manually verified, expanding its size requires significant resources and poses challenges. Nonetheless, the number of images and questions in Tri-HE is comparable to existing LVLM hallucination evaluation benchmarks such as Zhao et al. (2023) and Guan et al. (2024). Furthermore, as demonstrated in §5, Tri-HE is able to produce reliable hallucination evaluation results.

[4]We use the "gpt-4-vision-preview" model here, the same as in §5.2.
[5]Check §A.3 for the prompt used here.

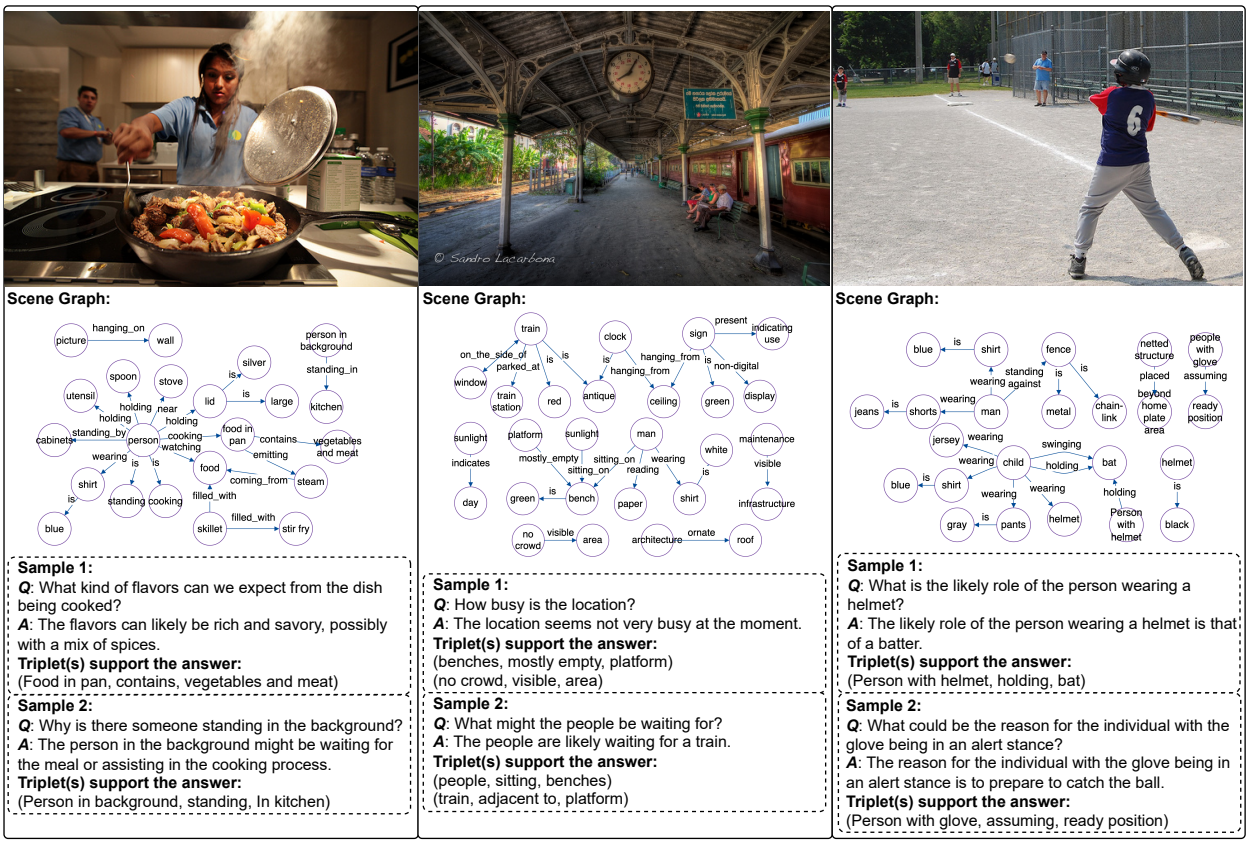


Figure 2: **Visualization of data samples in Tri-HE**. Each image is associated with a scene graph and question-answer pairs with the reasoning triplet annotations.

5 Evaluation Results

5.1 Evaluated LVLMs

We selected six open-source LVLMs for evaluation, including the LLaVA series (Liu et al., 2023a), MiniGPT-4 (Zhu et al., 2023), InstructBLIP (Dai et al., 2023), Shikra (Chen et al., 2023b), and InternLM-XComposer2 (*abbrev.*, InternLM2) (Cai et al., 2024). For all evaluated LVLMs, we selected the 7B variants to ensure fair comparison. Additionally, we test the recent popular Llama-3.2-Vision-Instruct model (*abbrev.*, LLaMA-3.2) (MetaAI, 2024a) and used its smallest version (11B). The prompt templates and inference configurations used for LVLMs are detailed in Appendix §A.4 and §C. All experiments are conducted on two Nvidia A100 GPUs.

5.2 Main Result

LVLM comparison. Table 2 compares hallucination rates of different LVLMs on our Tri-HE benchmark. As can be seen, all the evaluated LVLMs suffer from generating hallucinations with at least 38% hallucination rates. Among these LVLMs, InternLM2 (Cai et al., 2024) obtains the best overall performances, suggesting that its strategy to train with both text-image and textual-only instruction data simultaneously helps better align its visual encoder and LLM, and thus, reduces its hallucination rates. Moreover, compared to LLaVA (Liu et al., 2023b), Shikra (Chen et al., 2023b) has consistently lower hallucination rates, which is built upon LLaVA's structure with extra grounding capability introduced, indicating that introducing extra grounding could help LVLMs reduce hallucination. Additionally, LLaMA-3.2 achieves the lowest relation hallucination rates, suggesting that a strong textual backbone can help mitigate relation hallucination. However, it exhibits a weaker ability to accurately identify objects, impacting its object and overall hallucination rates. **Since**

Method	LLM Judge						NLI Judge	
	Overall		Object		Relation		Overall	
	$\text{Hallu}_\text{I} \downarrow$	$\text{Hallu}_\text{Q} \downarrow$	$\text{Hallu}_\text{I} \downarrow$	$\text{Hallu}_\text{Q} \downarrow$	$\text{Hallu}_\text{I} \downarrow$	$\text{Hallu}_\text{Q} \downarrow$	$\text{Hallu}_\text{I} \downarrow$	$\text{Hallu}_\text{Q} \downarrow$
MiniGPT-4	53.60	51.79	28.32	26.77	25.25	24.98	55.61	53.36
InstructBLIP	46.68	45.57	22.19	20.88	24.50	24.69	58.25	55.56
LLaVA	42.34	41.30	19.88	18.50	22.46	22.80	54.49	51.51
Shikra	42.20	41.76	18.55	17.54	23.65	24.22	56.46	53.98
LLaVA-1.5	40.66	39.10	18.63	**17.28**	22.03	21.82	54.14	51.67
LLaMA-3.2	40.16	38.95	22.30	21.08	**17.86**	**17.87**	**48.46**	**45.64**
InternLM2	**38.83**	**37.54**	**18.25**	17.50	20.58	20.04	54.41	52.08

Table 2: **Comparison on hallucination rates among different LVLMs on Tri-HE.** The best results under each column are **boldfaced**. InternLM2 is short for InternLM-XComposer2 (Cai et al., 2024)

Method	LLaVA	LLaVA-1.5	MiniGPT-4	InstructBLIP	Shikra	InternLM2	GPT-4V
NLI Judge (Sentence)	0.2182	0.0970	0.3609	0.2596	0.2684	0.2524	0.2787
NLI Judge (Triplet)	0.2951	0.2838	0.2264	0.4259	0.2829	0.2647	0.4190
LLM Judge (Llama-3.3 Sentence)	0.4705	0.4842	0.3617	0.2520	0.4941	0.4366	0.4969
LLM Judge (Llama-3.3 Triplet)	0.5138	0.5262	0.4150	0.4798	0.5311	0.5323	0.5519
LLM Judge (GPT-4 Sentence)	0.6631	0.5409	0.3669	0.5532	0.5821	0.5998	0.5548
LLM Judge (GPT-4 Triplet)	**0.8115**	**0.6320**	**0.4283**	**0.6235**	**0.6939**	**0.7169**	**0.7292**

Table 3: **Pearson correlation scores** among automatic hallucination judgments and human judgments. The best results under each column are **boldfaced**. The specific LLMs used in LLM Judge are specified in the brackets.

LLaMA-3.2 does not outperform other LVLMs with even more parameters, we do not adopt it in the remaining experiments for parameter consistency.

Relation hallucination is more severe. Except for MiniGPT-4 and LLaMA-3.2, all the LVLMs generate more relation hallucinations than object hallucinations. A possible explanation is that existing LVLMs lack reasoning abilities, which makes them easily confused and mess up the relations among objects. This further suggests that focusing on object hallucination (Li et al., 2023d) is not enough for a throughout analysis of the LVLM reliability, and a unified and comprehensive study like our proposed triplet-level evaluation is necessary.

Evaluation pipeline. In addition, we observe that LLM Judge can provide clearer and more reasonable discrimination between models compared to NLI judge. We provide a more comprehensive investigation into the differences between these two judges later in §5.3. Besides, the evaluation results under both Hallu_I and Hallu_Q metrics demonstrate the same trend, proving the robustness of our proposed triplet-level hallucination evaluation setting under different evaluation granularities.

Evaluate Close-sourced LVLMs. In addition to evaluating open-sourced LVLMs, we further investigate the performance of closed-sourced LVLMs. Due to limited experimental resources, we specifically evaluate the GPT-4V (OpenAI, 2023) model on a subset of 25 randomly selected images from Tri-HE. Specifically, we prompt GPT-4V to obtain responses to all questions related to these selected images and compute the associated hallucination rates following the steps described in Table 2. For comparison, we also include results from open-sourced LVLMs evaluated on the same set of 25 images. As illustrated in Figure 1b, GPT-4V clearly demonstrates superior performance, surpassing all open-sourced LVLMs. Although GPT-4V exhibits slightly higher object hallucination rates compared to InternLM2—likely because it tends to associate

		LLaVA	LLaVA-1.5	MiniGPT-4	InstructBLIP	Shikra	InternLM2
Original	$\text{Hallu}_\text{I} \downarrow$	22.46	22.03	25.25	24.50	23.65	20.58
	$\text{Hallu}_\text{Q} \downarrow$	22.80	21.82	24.98	24.69	24.22	20.04
First 20%	$\text{Hallu}_\text{I} \downarrow$	20.86	18.44	23.00	21.73	22.47	18.57
	$\text{Hallu}_\text{Q} \downarrow$	18.73	18.06	22.68	19.82	19.34	16.10

Table 4: **Relation hallucination rates for the top 20% frequent object pairs** of different LVLMs under the LLM Judge. **Original** refers to the results in Table 2.

additional objects not present in the image—it achieves notably lower relation hallucination rates due to its stronger reasoning capabilities, resulting in lower overall hallucination rates.

5.3 Analysis

Investigating automatic hallucination judgments with human judgments. In §3, we propose to measure hallucination on triplet-level and design two automatic hallucination judges. Here, we further illustrate the effectiveness of the triplet-level evaluation setting by studying its correlation with human judgments. To conduct fine-grained hallucination analysis, previous works (Jing et al., 2023; Min et al., 2023) split a model response into sub-sentences first, on which their hallucination measurements are conducted. We regard this method as a baseline for comparison. Specifically, we sample a subset of 20 images from Tri-HE and invite human annotators to score five-point-scale hallucination rates of the responses of all the LVLMs in §5.1 (check Appendix §D for the detailed annotation guidelines). The human annotators achieve a Krippendorff's alpha score of 0.66, indicating a high inter-agreement.

Results are shown in Table 3. We find that triplet-level hallucination rates have higher correlations with human judgments with both NLI and LLM Judges, indicating that identifying hallucination on triplets can lead to a more accurate, human-preferred evaluation for model responses. Moreover, we notice that the LLM Judges achieves a higher correlation to human judgments compared to the NLI counterpart, revealing LLMs' superior abilities to find hallucinations, which is also consistent with our observation in §5.2.

Applying Different LLMs in LLM Judge. While GPT-4 allows the LLM Judge to produce reliable hallucination evaluations, the associated API expenses could become large when evaluating a large number of examples. To mitigate potential cost constraints, we also examine whether alternative open-source LLMs can serve effectively in LLM Judge. Specifically, we replace GPT-4 with LLaMA-3.3-70B-Instruct (*abbrev.*, Llama-3.3) (MetaAI, 2024b) and re-evaluate all examples listed in Table 3. As shown, similar to GPT-4, Llama-3.3 consistently achieves higher correlation scores at the triplet-level than at the sentence-level. Furthermore, its Pearson correlation scores with human evaluations, while significantly outperforming those obtained using NLI Judge, remain comparable to GPT-4's results for certain LVLMs. These findings suggest that open-source LLMs can serve as viable alternatives to GPT-4 in LLM Judge, providing reliable evaluation results under tight budget constraints, thereby further validating the robustness of our proposed LLM Judge.

Investigating relation hallucination with object information. As concluded from §5.2, existing LVLMs tend to generate both object and relation hallucinations in their replies, while the relation hallucination rates are even higher. Since different LVLMs have pairs of objects (v_1, v_2) that they are familiar with (*e.g.*, high-frequency object pairs in the instruction data they are fine-tuned on) and might generate correct relations on these objects easily, we suppose that the relation hallucination problem might mostly be located in less-frequent object pairs. To verify this assumption, we extract all object pairs for each LVLM from their respective G_θ generated from responses on Tri-HE, and rank these pairs based on their frequency [6]. Then, we calculate each LVLM's relation hallucination rates on their most frequent object pairs.

[6]Details of this process can be seen in Appendix E.

As in Table 4, all the LVLMs have significantly lower relation hallucination rates on frequent object pairs they are familiar with, suggesting that they know the possible relations among objects and understand how to choose a relation appropriately.

Investigating hallucination rates with response length. Previous studies on LVLM hallucination evaluation suggest that the length of model responses may influence the extent of hallucination (Li et al., 2023d; Zhou et al., 2023), as some LVLMs tend to produce shorter, safer outputs. However, directly instructing an LVLM to generate a response of a specific length is challenging. To address this, we instead truncate the responses to the first K tokens and compute hallucination rates, varying K to assess its impact on the results.

As shown in Figure 3, while the exact hallucination rates vary, the ranking of different LVLMs remains consistent as the number of tokens increases from 10. Overall, as fewer tokens provide insufficient data for triplet extraction, this finding supports the robustness of our proposed triplet-level evaluation across LVLMs with varying response lengths.

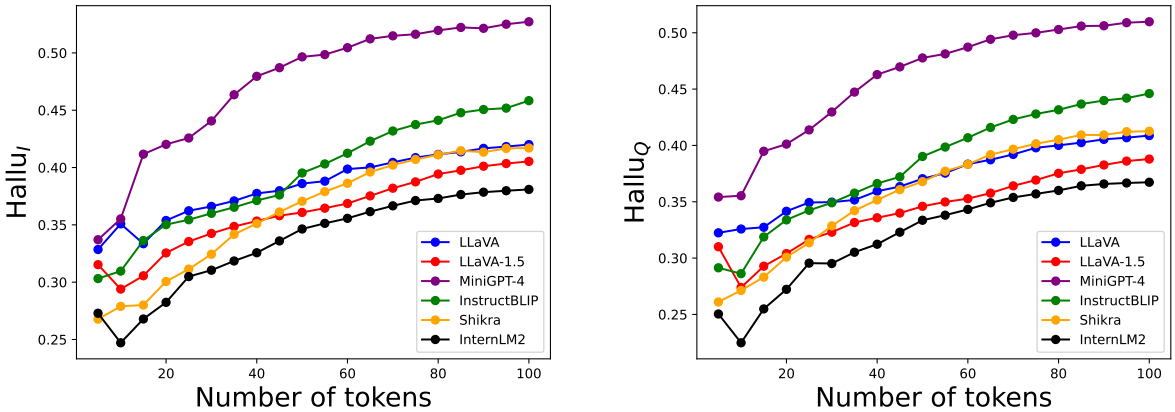


Figure 3: **Trends of the hallucination rates** of the image-level (left) and question-level (right) evaluations for different LVLMs with respect to the number of tokens in the model responses.

5.4 Hallucination Mitigation

After demonstrating that LVLMs exhibit significant hallucination problems, we further explore potential approaches to reduce both object and relation hallucinations. Prior works (Jing et al., 2023; Zhou et al., 2023; Li et al., 2023c; Gou et al., 2024) have suggested that *modality misalignment* might be a primary cause behind LVLM hallucinations. Motivated by this claim, we propose a training-free method to mitigate hallucinations by improving modality alignment within LVLMs.

Method. Specifically, we propose a two-step strategy. Given an image and its corresponding question, we first prompt the evaluated LVLM to generate a description of the image guided by the given question (**General Description** in Figure 4). Next, we prompt the same LVLM (in a new version without image memory) using this generated description to answer the question. Through this approach, we effectively leverage the strong instruction-following capability intrinsic to the LVLM's LLM backbone, instead of requiring the LVLM to simultaneously comprehend the image and answer the question, thereby reducing hallucinations caused by modality misalignment. Moreover, as indicated in §5.3, triplet-level evaluation is more effective than sentence-level evaluation in assessing hallucinations. Hence, we further explicitly guide LVLMs to concentrate more on identifying objects and their interrelations in the form of triplets when describing images (**Triplet Description** in Figure 4).

We evaluate MiniGPT-4 and LLaVA-1.5 combined with our proposed mitigation approaches using the subset previously employed in Figure 1b. The corresponding prompts along with an example illustration are shown in Figure 4.

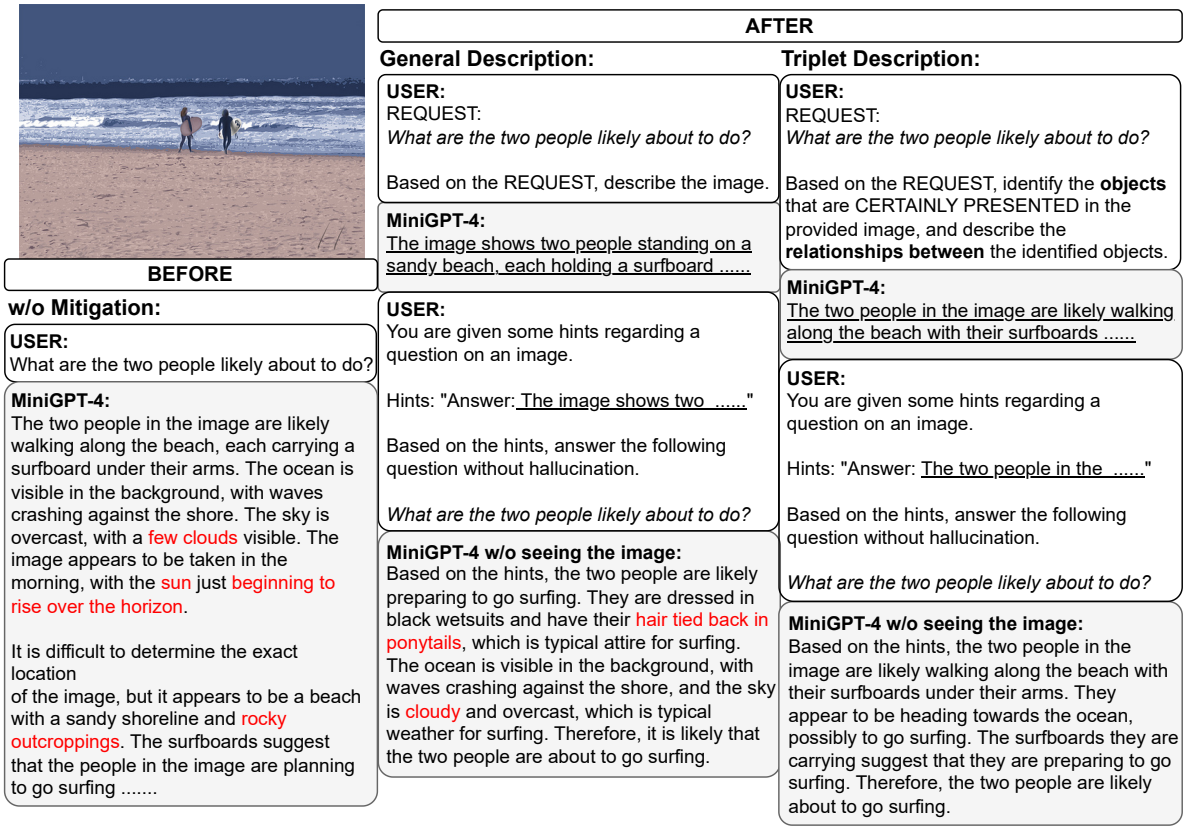


Figure 4: **An illustration demonstrating hallucination mitigation**. The three prompting strategies (w/o Mitigation, General Description, and Triplet Description) are listed from left to right. Hallucinated content is highlighted in Red and repeating contents are marked with *italic* and underline.

Results. As demonstrated in Table 5, both LVLMs exhibit reduced hallucination rates after applying our two-stage mitigation method, indicating that improved modality alignment effectively alleviates hallucinations. In addition, explicitly prompting LVLMs to emphasize objects and their relationships consistently yields the lowest hallucination rates across most cases, further reinforcing the findings presented in Table 3.

Ablation Study. We also perform an ablation study on the best-performing "Triplet Description" variant of our mitigation approach to gain deeper insights into the role of each module within our proposed method. Specifically, we compare the Triplet Description (*i.e., Triplet+Eyes-Close*) results obtained by MiniGPT-4 with two alternative setups:

1. *Eyes-Close:* This setting is equivalent to General Description. Image access is disabled (i.e., eyes-close) while prompting LVLM to answer the question. It is designed to assess the impact of employing triplet-level descriptions.

2. *Triplet:* This setting is similar to Triplet Description but allows image accessibility. It incorporates both the original image and the generated triplet-level description simultaneously as inputs. It is designed to examine the effects of modality alignment.

The experimental results are presented in Table 6. As shown, the combined use of triplet-level description and restriction of visual input access leads to the lowest hallucination rates. These findings further validate the design choices made in our mitigation method.

	Mitigation	LLM Judge		NLI Judge	
		$\text{Hallu}_\text{I} \downarrow$	$\text{Hallu}_\text{Q} \downarrow$	$\text{Hallu}_\text{I} \downarrow$	$\text{Hallu}_\text{Q} \downarrow$
MiniGPT-4	w/o Mitigation	45.86	47.44	55.93	54.94
	General Description	46.50	49.19	54.59	53.03
	Triplet Description	**44.14**	**42.96**	**51.19**	**47.12**
LLaVA-1.5	w/o Mitigation	30.72	30.17	53.84	52.06
	General Description	28.70	**29.80**	51.40	49.80
	Triplet Description	**28.39**	32.68	**48.97**	**48.40**

Table 5: **Hallucination mitigation** results. The best results under each column are **boldfaced**.

Mitigation	LLM Judge		NLI Judge	
	$\text{Hallu}_\text{I} \downarrow$	$\text{Hallu}_\text{Q} \downarrow$	$\text{Hallu}_\text{I} \downarrow$	$\text{Hallu}_\text{Q} \downarrow$
w/o Mitigation	45.86	47.44	55.93	54.94
Eyes-Close	46.50	49.19	54.59	53.03
Triplet	45.65	45.16	59.35	55.57
Triplet+ Eyes-Close	**44.14**	**42.96**	**51.19**	**47.12**

Table 6: **Ablation study on MiniGPT-4 (Zhu et al., 2023).** The best results under each column are **boldfaced**.

6 Conclusion

Starting from a unified definition of hallucinations, we propose a novel triplet-level LVLM hallucination evaluation framework for both object and relation hallucinations. Then we introduce Tri-HE, a novel triplet-level LVLM hallucination evaluation benchmark, with which, we conduct a throughout analysis of the discrepancy among object and relation hallucinations. Finally, we propose a simple yet effective training-free hallucination mitigation method, which integrates our findings regarding objects and inter-object relations.

References

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. Internlm2 technical report. *ArXiv preprint*, abs/2403.17297, 2024. URL `https://arxiv.org/abs/2403.17297`.

Kai Chen, Chunwei Wang, Kuo Yang, Jianhua Han, Lanqing Hong, Fei Mi, Hang Xu, Zhengying Liu, Wenyong Huang, Zhenguo Li, Dit-Yan Yeung, Lifeng Shang, Xin Jiang, and Qun Liu. Gaining wisdom from setbacks: Aligning large language models via mistake analysis. *ArXiv preprint*, abs/2310.10477, 2023a. URL `https://arxiv.org/abs/2310.10477`.

Kai Chen, Yunhao Gou, Runhui Huang, Zhili Liu, Daxin Tan, Jing Xu, Chunwei Wang, Yi Zhu, Yihan Zeng, Kuo Yang, et al. Emova: Empowering language models to see, hear and speak with vivid emotions. *arXiv preprint arXiv:2409.18042*, 2024a.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *ArXiv preprint*, abs/2306.15195, 2023b. URL https://arxiv.org/abs/2306.15195.

Xiang Chen, Chenxi Wang, Yida Xue, Ningyu Zhang, Xiaoyan Yang, Qiang Li, Yue Shen, Jinjie Gu, and Huajun Chen. Unified hallucination detection for multimodal large language models. *ArXiv preprint*, abs/2402.03190, 2024b. URL https://arxiv.org/abs/2402.03190.

Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Shengyi Qian, Jianing Yang, David F Fouhey, and Joyce Chai. Multi-object hallucination in vision-language models. *arXiv preprint arXiv:2407.06192*, 2024c.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. Halc: Object hallucination reduction via adaptive focal-contrast decoding. *ArXiv preprint*, abs/2403.00425, 2024d. URL https://arxiv.org/abs/2403.00425.

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

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *ArXiv preprint*, abs/2401.16420, 2024. URL https://arxiv.org/abs/2401.16420.

Yunhao Gou, Zhili Liu, Kai Chen, Lanqing Hong, Hang Xu, Aoxue Li, Dit-Yan Yeung, James T Kwok, and Yu Zhang. Mixture of cluster-conditional lora experts for vision-language instruction tuning. *ArXiv preprint*, abs/2312.12379, 2023. URL https://arxiv.org/abs/2312.12379.

Yunhao Gou, Kai Chen, Zhili Liu, Lanqing Hong, Hang Xu, Zhenguo Li, Dit-Yan Yeung, James T Kwok, and Yu Zhang. Eyes closed, safety on: Protecting multimodal llms via image-to-text transformation. *ArXiv preprint*, abs/2403.09572, 2024. URL https://arxiv.org/abs/2403.09572.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. 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*, pp. 14375–14385, 2024.

Zongbo Han, Zechen Bai, Haiyang Mei, Qianli Xu, Changqing Zhang, and Mike Zheng Shou. Skip \n: A simple method to reduce hallucination in large vision-language models. *ArXiv preprint*, abs/2402.01345, 2024. URL https://arxiv.org/abs/2402.01345.

Qidong Huang, Xiaoyi Dong, Pan zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *ArXiv preprint*, abs/2311.17911, 2023. URL https://arxiv.org/abs/2311.17911.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 6700–6709. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00686. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Hudson_GQA_A_New_Dataset_for_Real-World_Visual_Reasoning_and_Compositional_CVPR_2019_paper.html.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. In *ACM Computing Surveys*, 2023.

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

Klaus Krippendorff. Computing krippendorff's alpha-reliability, 2011.

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

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *EMNLP*, 2023b.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6449–6464, Singapore, 2023c. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.397. URL https://aclanthology.org/2023.emnlp-main.397.

Yanze Li, Wenhua Zhang, Kai Chen, Yanxin Liu, Pengxiang Li, Ruiyuan Gao, Lanqing Hong, Meng Tian, Xinhai Zhao, Zhenguo Li, et al. Automated evaluation of large vision-language models on self-driving corner cases. *ArXiv preprint*, abs/2404.10595, 2024. URL https://arxiv.org/abs/2404.10595.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, 2023d.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *ArXiv preprint*, abs/2310.03744, 2023a. URL https://arxiv.org/abs/2310.03744.

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

MetaAI. Introducing llama 3.2, 2024a. URL https://www.llama.com/.

MetaAI. Llama-3.3-70b-instruct, 2024b. URL https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct.

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

OpenAI. ChatGPT, 2023. URL https://openai.com/research/gpt-4v-system-card.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410.

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

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *ArXiv preprint*, abs/2311.07397, 2023b. URL https://arxiv.org/abs/2311.07397.

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. Evaluation and analysis of hallucination in large vision-language models. *ArXiv preprint*, abs/2308.15126, 2023c. URL https://arxiv.org/abs/2308.15126.

Mingrui Wu, Jiayi Ji, Oucheng Huang, Jiale Li, Yuhang Wu, Xiaoshuai Sun, and Rongrong Ji. Evaluating and analyzing relationship hallucinations in lvlms. *arXiv preprint arXiv:2406.16449*, 2024.

Zhao Xiaoyan, Deng Yang, Yang Min, Wang Lingzhi, Zhang Rui, Cheng Hong, Lam Wai, Shen Ying, and Xu Ruifeng. A comprehensive survey on deep learning for relation extraction: Recent advances and new frontiers. *ArXiv preprint*, abs/2306.02051, 2023. URL `https://arxiv.org/abs/2306.02051`.

Zihao Yue, Liang Zhang, and Qin Jin. Less is more: Mitigating multimodal hallucination from an eos decision perspective. *ArXiv preprint*, abs/2402.14545, 2024. URL `https://arxiv.org/abs/2402.14545`.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*, 2023.

Kening Zheng, Junkai Chen, Yibo Yan, Xin Zou, and Xuming Hu. Reefknot: A comprehensive benchmark for relation hallucination evaluation, analysis and mitigation in multimodal large language models. *arXiv preprint arXiv:2408.09429*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *ArXiv preprint*, abs/2306.05685, 2023. URL `https://arxiv.org/abs/2306.05685`.

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

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *ArXiv preprint*, abs/2304.10592, 2023. URL `https://arxiv.org/abs/2304.10592`.

A Prompts

A.1 Prompt for triplets extraction with GPT-4

The prompt for extracting triplets in the answer generated by LVLMs is illustrated in Figure 5.

Given a description of the image, please extract a KG from the text and represent the KG with triples formatted with ("subject", "predicate", "object") with readability, each triplet in a line. If 'and' or 'or' exists in the input sentence, split the objects into multiple triplets. Please do not extract from uninformative sentences.

Here are some in-context examples:

Input:
Optimus (or Tesla Bot) is a robotic humanoid under development by Tesla, Inc. It was announced at the company's Artificial Intelligence (AI) Day event on August 19, 2021. It is planned to measure 5 ft 8 in (173 cm) tall and weigh 125 lb (57 kg). It is hard to answer whether Tesla is good to invest without more information.

KG:
("Optimus", "is", "robotic humanoid")
("Optimus", "under development by", "Tesla, Inc.")
("Optimus", "also known as", "Tesla Bot")
("Tesla, Inc.", "announced", "Optimus")
("Announcement of Optimus", "occured at", "Artificial Intelligence (AI) Day event")
("Artificial Intelligence (AI) Day event", "held on", "August 19, 2021")
("Artificial Intelligence (AI) Day event", "organized by", "Tesla, Inc.")
("Optimus", "planned to measure", "5 ft 8 in (173 cm) tall")
("Optimus", "planned to measure", "weigh 125 lb (57 kg).")
<Done>

Input:
The image doesn't provide information about the popularity of the song. The song "Here Comes the Boom" was originally released by American rock band Nelly in 2002 for the soundtrack of the film "The Longest Yard."

KG:
("The song 'Here Comes the Boom'", "originally released by", "American rock band Nelly")
("The song 'Here Comes the Boom'", "released in", "2002")
("The song 'Here Comes the Boom'", "featured in", "soundtrack of the film 'The Longest Yard'")
("American rock band Nelly", "released", "The song 'Here Comes the Boom'")
("The Longest Yard", "had soundtrack featuring", "The song 'Here Comes the Boom'")
<Done>

Now generate the KG for the provided input text:

Input:
{input_text}

KG:

Figure 5: **Prompt for triplets extraction with GPT-4.**

A.2 Prompt for LLM Judge

The prompt for our proposed LLM judge method is illustrated in Figure 6.

Given a list of reference triplets ("object1", "relation", "object2") extracted from the scene graph of an image, along with a list of objects observed in this image, your task is:

Task 1. Determine if a claim triplet ("object1", "relation", "object2") is directly supported by any single triplet in the reference, or can be logically inferred from multiple reference triplets and the list of objects. Follow these steps when finishing the task:

1. Answer "yes" if the claim appears in the reference.

2. Answer "yes" if the claim can be logically inferred from one or more triplets in the reference. Consider:

a. General Inferences: Assess common associations or implications.
b. Conditional Phrases: Note phrases like "could be", "might", "suggests", which allow broader inferences.
c. Equivalence of Objects: In your judgment, treat objects of the same kind as equal. For example, "woman", "man" should be considered under the general category of "person".
d. Support from Object List: If the claim is not directly supported or inferable from the triplets, assess whether the list of objects provides additional evidence to support or infer the claim.

3. Answer "no" if the claim neither directly matches any triplet in the reference nor can be reasonably inferred from the triplets and the object list.

Task 2: Error categorization.

If your answer to the previous task is "no", determine whether the not supported/inferred part in the claim is "object1" or "object2" or "relation".

Reference:
<REFERENCE>

List of Objects:
<LIST_OF_OBJECTS>

Claim:
<CLAIM>

Figure 6: **Prompt for the LLM Judge method.**

A.3 Prompt for question generation with GPT-4V

The prompt for generating questions, answers, and corresponding triplets with GPT-4V is shown in Figure 7.

A.4 Prompts for Evaluating LVLMs

When evaluating LVLMs on Tri-HE, the prompt we use is the question itself. Questions are fed into LVLMs along with the corresponding images.

Generate ten questions about the given image that require an inferential answer, which is not directly observable from the image. The answer to each question can be explained by one or more (object, relation, object) triplets that appear in the scene graph of the given image. Note that the triplets should consist of objects and relations that are visible in the given image. Output the results in the format of:

Generated Questions:

Answers:

Explanations:

Figure 7: **Prompt for question generation with GPT-4V.**

B NLI Threshold Selection

We randomly selected question instances from 10 images and reviewed the set of filtered triplets that were returned. The similarity score threshold was adjusted to 0.5 for the most reasonable returned triplets. These triplets later concatenate together as the ground truth required for generating NLI judgments. In determining if a generated triplet was hallucinated, we further review the NLI judgment results in different thresholds, ultimately deciding on a threshold of 0.6.

C Configurations for LVLM Evaluation

For LVLM evaluations, we directly use the default configuration settings provided in their publicly available code repositories. For instance, the configurations utilized for evaluating LLaVA models are accessible at `https://github.com/haotian-liu/LLaVA`.

D Human Annotation Guideline

The detailed guidelines of our human evaluation tasks are shown in Table 7. Noting that two types of inferences in the model responses are regarded as hallucinations during human annotation:

1. Unreasonable inferences (inferences that violate commonsense knowledge).
2. Inferences that are correct, yet cannot be correctly inferred from the image.

Score	Description
1	1) The text is totally hallucinated, and is irrelevant to the given input image and question. or 2) The text is very hard to understand.
2	1) Most of the given responses are hallucinated, yet few sentences of them (one or two) are related to the given image and question.
3	1) Half of the sentences in the given response are hallucinated.
4	1) Most of the sentences in the generated response are not hallucinated.
5	1) No hallucination exists in the generated response.

Table 7: **Detailed human evaluation instructions.**

E Object Pairs Extraction and Ranking

In this section, we detailedly describe how we obtain object pairs and their rankings from LVLM responses. Suppose we have all an LVLM's responses to all questions in Tri-HE, *i.e.*, G_θ, we first extract all the object pairs (v_1, v_2) from G_θ. Then for each object, we replace it by the name of its synset using WordNet to reduce the total types of objects. Afterward, we could calculate the frequency of each object pair and rank them based on their frequency. This ranking will then be used to calculate the first 20% frequent object pairs in Table 4.