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

Junjie Wu* **Tsz Ting Chung*** **Kai Chen*** **Dit-Yan Yeung**
The Hong Kong University of Science and Technology
{junjie.wu, ttchungac, kai.chen}@connect.ust.hk dyyeung@cse.ust.hk
https://kaichen1998.github.io/projects/tri-he/

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, in this paper we design a unified framework to measure object and relation hallucination in LVLMs simultaneously. The core idea of our framework is to conduct hallucination evaluation on (object, relation, object) triplets extracted from LVLMs' responses, and thus, could be easily generalized to different 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. We conduct comprehensive evaluations on Tri-HE and 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 yet effective training-free approach to mitigate hallucinations for LVLMs, with which, we exceed all open-sourced counterparts on Tri-HE, achieving comparable performance with the powerful GPT-4V. Our dataset and code for the reproduction of our experiments are available publicly at `https://github.com/wujunjie1998/Tri-HE`.

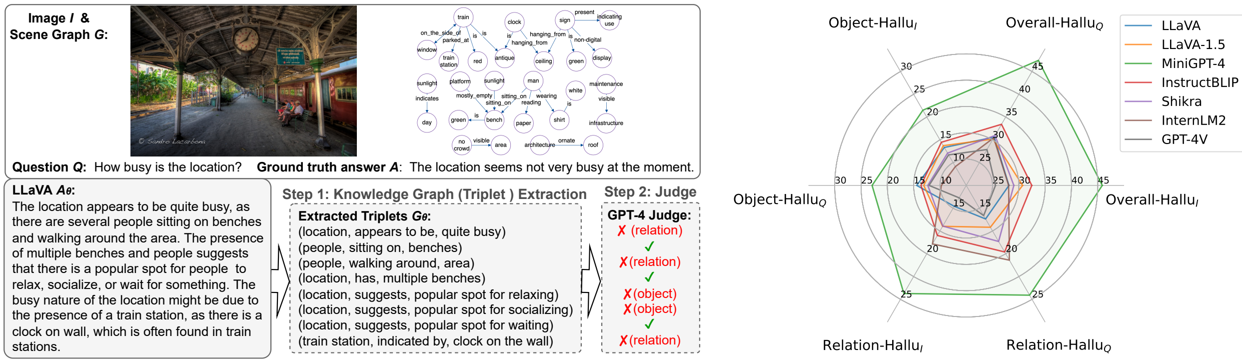
1 Introduction

Large Vision-Language Models (LVLMs) [1, 2, 3, 4] 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 [5, 6, 7, 8, 9], *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 [2] 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* [5, 6, 10] (*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:** existing works [5, 6, 8, 11] rely on self-discrimination (*e.g.*, Yes/No questions) to testify LVLMs' awareness of objects and relations in the given image. When prompted with an instruction *"Is there a*

[1]* Equal contribution



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

Figure 1: **Overview of Tri-HE 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 GPT-4. (b) The radar plot showcases the evaluation results among different LVLMs (lower values demonstrate fewer hallucinations).

{content} in the image?", hallucination is determinied only when LVLMs answer *"Yes."* for a non-existing content. However, these benchmarks require transforming general vision-language tasks into "yes/no" formats, limiting their applicability. Additionally, different LVLMs may have different capability in answering such"yes/no" questions since they are pre-trained on different data, which may make the evaluation results biased [12]. So, here arises our question: **can we perform a unified and unbiased evaluation for various kinds of hallucination?**

To this end, this paper first proposes a unified framework to measure both object and relation hallucinations (§3). Specifically, it first extracts knowledge graphs as triplets from LVLMs' responses, then uses external judges to compare these triplets with scene graphs of input images for hallucination evaluation. In this way, our method can provide fine-grained evaluation results on the response to any vison-language tasks. Based on the proposed framework, we then construct **Tri-HE**, a novel **Tri**plet-level **H**allucination **E**valuation benchmark to evaluate object and relation hallucination simultaneously (§4). As reported in §5 and Figure 1b, relation hallucination is indeed a common nightmare even for the powerful GPT-4V model [13], and is even more severe than object hallucination. By comparing the performance of different LVLMs, we further recognize several findings that could potentially decrease LVLM hallucination (§5.2). In addition, our triplet-level hallucination judge empowered by GPT-4 reveals remarkable alignment with human judgments (Table 3). This motivates us to integrate triplet description into LVLMs and propose a simple yet effective training-free hallucination mitigation method (§5.4). With this method, the LLaVA-1.5 model [14] exceeds all the open-sourced counterparts, achieving comparable performance with the powerful GPT-4V, as shown in Figure 1b and Table 5.

Our main contributions contain three parts:

1. We introduce a unified framework to measure both the object and relation hallucination of LVLMs simultaneously. Specifically, we propose to conduct a triplet-level hallucination evaluation, which can provide more accurate and fine-grained evaluation results.
2. Based on the framework, we propose **Tri-HE**, a novel triplet-level fine-grained hallucination evaluation benchmark for LVLMs.
3. We propose a simple yet effective training-free hallucination mitigation method that exceeds all open-sourced counterparts, performing on par with GPT-4V on Tri-HE.

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 [14, 2] proposes to use a simple projection layer to integrate the visual representations into textual encoders, which is further enhanced in Shikra [15] by incorporating referential dialogue tasks. Instead, BLIP [16] proposes the Q-Former architecture to extract useful information from the visual representations, which is also used by MiniGPT-4 [3] and

InstructBLIP [1]. MoCLE [17] 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 on 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 [9, 18, 19, 20, 21, 6, 8, 22]. Previous hallucination evaluation works can be categorized into two groups: 1) solely evaluating object hallucinations or do not distinguish different hallucinations [23, 24, 6, 10], 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 [5, 25, 8, 11]. 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 worth noting that a concurrent benchmark, Reefknot [26], also evaluates relation hallucinations at the triplet level. However, compared to Tri-HE, Reefknot has several limitations: 1) It fully relies on existing datasets, which may lead to data contamination, whereas Tri-HE consists of questions synthesized by GPT-4V. 2) Reefknot evaluates only a limited number of relation types, while Tri-HE covers a broader range, enabling more comprehensive evaluation results as shown in §5. 3) Reefknot employs only one entailment-based hallucination discriminator, whereas Tri-HE offers more accurate hallucination detection using GPT-4, which can simultaneously identify both object and relation hallucinations.

3 Tri-HE Framework Formulation

Inspired by the relation extraction [27] 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 shown in Figure 1a, given an input image I, a question Q related to I with its ground truth answer A, and the answer A_θ predicted by a LVLM $A_\theta(\cdot|Q, I)$ parameterized with 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** [28]. 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 hallucination rates of A_θ as the **proportion of hallucinated triplets** in G_θ. Most previous works (*e.g.*, POPE [5]) 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, **$\text{Hallu}_\text{Q}(\cdot)$ and $\text{Hallu}_\text{I}(\cdot)$ 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 Figure 6. Afterward, 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) [29] 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, suggesting the extracted triplet cannot be induced based on GT triplets, and therefore, resulting in a hallucination.

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

Table 1: **Statistics of Tri-HE.**

GPT-4 Judge. The other strategy is via prompting the GPT-4 model, which is a common practice for recent LLM evaluation literature [30]. We prompt GPT-4 to justify whether a previously extracted triplet $(v_1, e, v_2) \in G_\theta$ can be **directly obtained** or **inferred** based on the image scene graph G. Note that we do not adopt the GPT-4V judge, since [31] have observed that the text-only GPT-4 judge better aligns with the human preference than GPT-4V. Moreover, if (v_1, e, v_2) is recognized as a hallucinated triplet, we further require it to tell whether the hallucination relates to the relation e or objects v_1, v_2. Check our prompt for GPT-4 judge in Figure 8 [2].

With triplet-level hallucination judgments in hand, we can adopt Eq. (1) and Eq. (2) to calculate the corresponding hallucination rates of A_θ, which can be further separated into object hallucination and relation hallucination rates if GPT-4 judge is adopted. Check more detailed results in Table 2.

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

[2]For both knowledge graph extraction and GPT-4 judge, we use the "gpt-4-1106-preview" model via OpenAI's API calls with its default inference parameters.

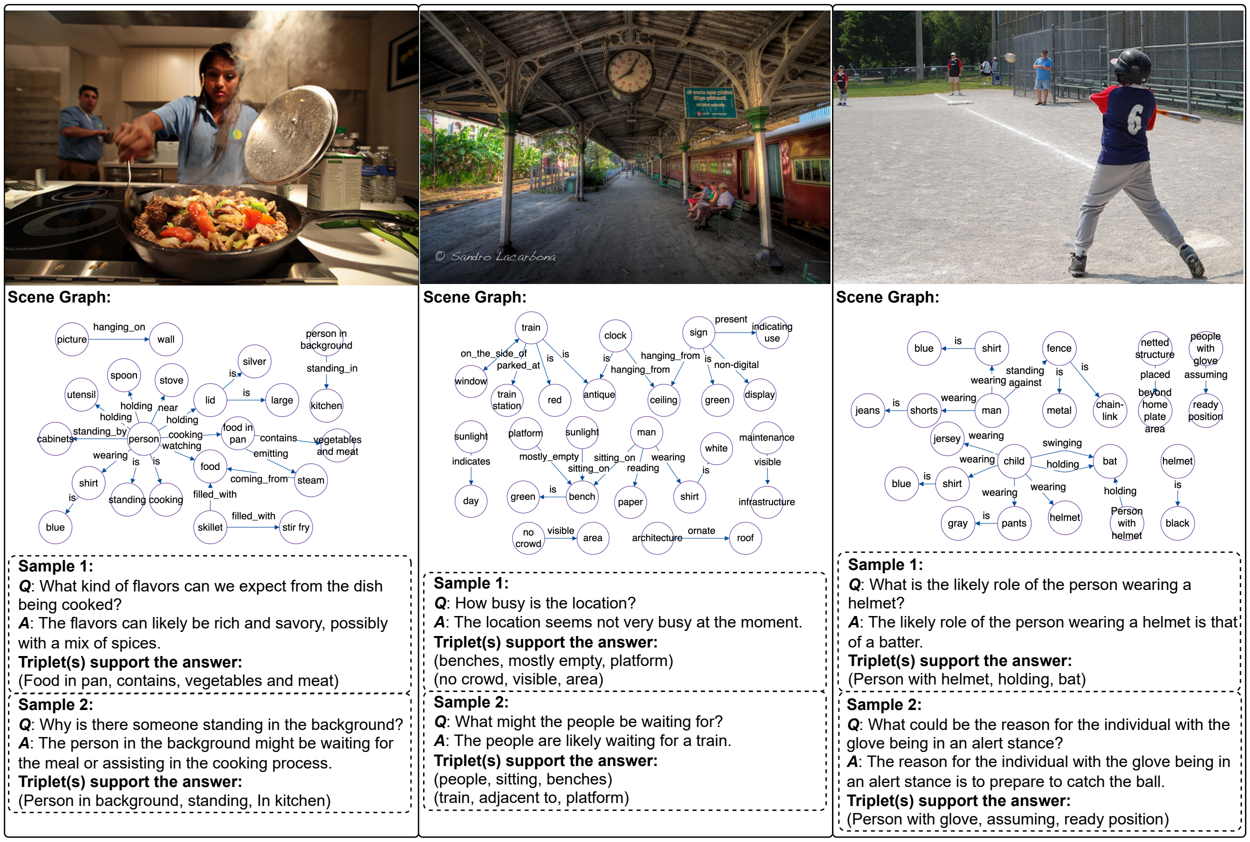


Figure 2: **Visualization of data samples in Tri-HE**. Each image is associated with a scene graph and question-answer pairs with the reasoning triplet annotations.

4 Tri-HE Construction

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

Dataset. We build Tri-HE upon the GQA dataset [32], which requires LVLMs to conduct commonsense reasoning based on a given image. The provision of scene graph annotations aligns with our formulation of hallucination evaluation at the triplet-level.

Data collection. However, several scene graphs in GQA are only equipped with parts of the relations among objects appearing in the image, making necessary information required for question answering absent. To address that, we first employ a filtering process to identify images with adequate distinct relations in the scene graph annotation. Specifically, we start by selecting 300 images with more than 5 relations from the corresponding GQA scene graph annotations. To enlarge the question-answer pairs for each GQA image, we utilize GPT-4V [3] to generate new questions requiring commonsense reasoning based on the input image. By leveraging the prompt provided in Figure 7, we prompt GPT-4V to generate ten questions for each image along with their corresponding answers and relation triplets that illustrate the reasoning process. Afterward, we manually verify the correctness of generated questions, answers, and triplets, while only questions indeed requiring reasoning are maintained. Moreover, the verified triplets are added to the original scene graphs collected in GQA.

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

[3]We use the "gpt-4-vision-preview" model here, as in §5.2.

Method	GPT-4 Judge						NLI Judge	
	Overall		Object		Relation		Overall	
	Hallu_I ↓	Hallu_Q ↓	Hallu_I ↓	Hallu_Q ↓	Hallu_I ↓	Hallu_Q ↓	Hallu_I ↓	Hallu_Q ↓
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 [33]

comparable to existing LVLM hallucination evaluation benchmarks such as [23] and [8]. Furthermore, as demonstrated in §5, Tri-HE is able to produce reliable hallucination evaluation results.

5 Results on the Tri-HE Benchmark

5.1 Evaluated LVLMs

We selected six open-source LVLMs for evaluation, including the LLaVA series [14, 2], MiniGPT-4 [3], InstructBLIP [1], Shikra [15], and InternLM-XComposer2 (*abbrev.*, InternLM2) [33]. 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) [34] and used its smallest version (11B). The official prompt templates and inference configurations were used for all LVLMs. 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 [33] 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 [2], Shikra [15] 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 LLaMA-3.2 does not outperform other LVLMs with even more parameters, we do not adopt it in the remaining.*

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 [5] 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 GPT-4 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.

Performance of GPT-4V. Except for the open-sourced LVLMs, we also explore the performance of state-of-the-art commercial LVLMs [13, 35]. Due to the experiment budget, here we only evaluate the GPT-4V [13] model with 25 randomly selected samples in Tri-HE. Specifically, we directly prompting GPT-4V to gather its answers and calculate the corresponding hallucination rates, following the same setting in Table 2. We also report the performances of open-sourced LVLMs on the 25 selected images. As can be seen in Figure 1b, GPT-4V demonstrates its superior

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
GPT-4 Judge (Sentence)	0.6631	0.5409	0.3669	0.5532	0.5821	0.5998	0.5548
GPT-4 Judge (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**.

		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 GPT-4 judge. **Original** refers to the results in Table 2.

capability and outperforms all open-sourced LVLMs. Although its object hallucination rates are slightly higher than InternLM2 since it tends to associate many objects that are not in the image, GPT-4V has significantly lower reasoning hallucination rates than other LVLMs, indicating its strong reasoning ability, and therefore, reports superior 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 [36, 37] 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 Table 9 for the detailed annotation guidelines. The human annotators achieve a Krippendorff's alpha score [38] 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 GPT-4 judges, indicating that identifying hallucination on triplets can lead to a more accurate, human-preferred evaluation for model responses. Moreover, we notice that the GPT-4 judge achieves a higher correlation to human judgments compared to the NLI counterpart, revealing GPT-4's superior ability to find hallucinations, which is also consistent with our observation in §5.2.

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 [4]. Then, we calculate each LVLM's relation hallucination rates on their most frequent object pairs.

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. Check more analyses regarding the impact of input length on hallucination rates in Appendix A.

5.4 Hallucination Mitigation

Regarding the reasons of LVLM hallucination, previous studies [36, 39] have identified *modality misalignment* as a key factor. Inspired by ECSO [40], we propose a training-free LVLM hallucination mitigation method via self-alignment [12, 40, 41]. Specifically, we disable the direct visual access of LVLMs to alleviate hallucinations resulting

[4] Details of this process can be seen in Appendix D.

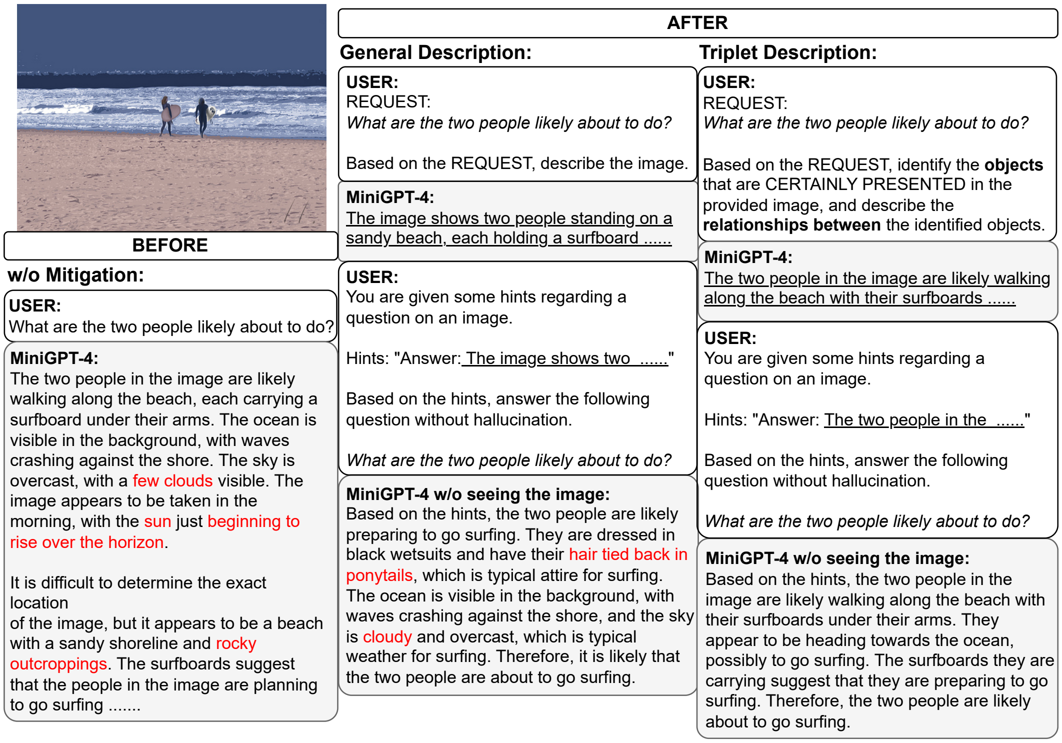


Figure 3: **An illustration demonstrating hallucination mitigation**. The three prompting strategies (w/o Mitigation, General Description, and Triplet Description) are listed from left to right. Hallucinated content is highlighted in Red and repeating contents are marked with *italic* and underline.

Mitigation	GPT-4 Judge		NLI Judge	
	Hallu_{I} ↓	Hallu_{Q} ↓	Hallu_{I} ↓	Hallu_{Q} ↓
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**

Table 5: **Hallucination mitigation** results on MiniGPT-4. The best results under each column are **boldfaced**.

from modality misalignment and fully leverage instruction-following capabilities of LVLMs' (text-only) LLM backbone for hallucinations mitigation [24].

Method. We start by prompting LVLMs to describe the image on awareness of the given question (*i.e.*, **General Description**). Subsequently, we prevent LVLMs from having direct access to the image to downgrade the text-image LVLMs to text-only LLMs, which are then encouraged to respond only based on the previously generated image description but without access to the image.

As discussed in §5.3, better hallucination evaluation is observed at the triplet level. Thus, we further prompt LVLMs to focus more on identifying objects and inter-object relations (*i.e.*, **Triplet Description**) as triplets when describing the image. We use the subset in Figure 1b with MiniGPT-4, with the prompts and an example in Figure 3.

Results. As shown in Table 5, our results demonstrate the effectiveness of hallucination mitigation through our proposed prompting strategies. Both general and triplet descriptions achieve a significant hallucination rate reduction while prompting LVLMs to focus on objects and inter-object relations achieves the lowest hallucination rates. Additionally, we conduct ablation studies and compare our method with existing baselines in Appendix B.

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.

Limitations

Due to the experiment budget, in Figure 1b, we evaluate GPT-4V on a randomly selected subset of 25 images; in Table 3, we conduct human annotation on another randomly selected subset of 20 images. Although we do not interfere with the data selection process, there may still exist a few biases in the corresponding results. Also, we do not have the budget to evaluate more LVLMs on Tri-HE. We will try to provide more comprehensive results in future works once we get more experimental budgets.

References

[1] 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.

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

[3] 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.

[4] 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*, 2024.

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

[6] 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, 2023.

[7] 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, 2023.

[8] 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*, pages 14375–14385, 2024.

[9] 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, 2024.

[10] 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*, 2024.

[11] 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.

[12] 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, 2023.

[13] OpenAI. ChatGPT, 2023.

[14] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *ArXiv preprint*, abs/2310.03744, 2023.

[15] 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, 2023.

[16] 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*, 2023.

[17] 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.

[18] 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, 2024.

[19] 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.

[20] 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.

[21] 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*, 2023.

[22] Zihao Yue, Liang Zhang, and Qin Jin. Less is more: Mitigating multimodal hallucination from an eos decision perspective. *ArXiv preprint*, abs/2402.14545, 2024.

[23] 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.

[24] 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, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore, 2023. Association for Computational Linguistics.

[25] 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*, 2023.

[26] 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.

[27] 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.

[28] 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.

[29] 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)*, pages 3982–3992, Hong Kong, China, 2019. Association for Computational Linguistics.

[30] 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.

[31] 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.

[32] 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*, pages 6700–6709. Computer Vision Foundation / IEEE, 2019.

[33] 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.

[34] MetaAI. Introducing llama 3.2, 2024.

[35] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *ArXiv preprint*, abs/2312.11805, 2023.

[36] 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.

[37] 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.

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

[39] 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.

[40] 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.

[41] Zhili Liu, Yunhao Gou, Kai Chen, Lanqing Hong, Jiahui Gao, Fei Mi, Yu Zhang, Zhenguo Li, Xin Jiang, Qun Liu, et al. Mixture of insightful experts (mote): The synergy of thought chains and expert mixtures in self-alignment. *ArXiv preprint*, abs/2405.00557, 2024.

[42] Junfei Wu, Qiang Liu, Ding Wang, Jinghao Zhang, Shu Wu, Liang Wang, and Tieniu Tan. Logical closed loop: Uncovering object hallucinations in large vision-language models. *arXiv preprint arXiv:2402.11622*, 2024.

[43] Kai Chen, Enze Xie, Zhe Chen, Lanqing Hong, Zhenguo Li, and Dit-Yan Yeung. Integrating geometric control into text-to-image diffusion models for high-quality detection data generation via text prompt. *ArXiv preprint*, abs/2306.04607, 2023.

[44] Ruiyuan Gao, Kai Chen, Zhihao Li, Lanqing Hong, Zhenguo Li, and Qiang Xu. Magicdrive3d: Controllable 3d generation for any-view rendering in street scenes. *ArXiv preprint*, abs/2405.14475, 2024.

[45] Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. Magicdrive: Street view generation with diverse 3d geometry control. *ArXiv preprint*, abs/2310.02601, 2023.

[46] Pengxiang Li, Zhili Liu, Kai Chen, Lanqing Hong, Yunzhi Zhuge, Dit-Yan Yeung, Huchuan Lu, and Xu Jia. Trackdiffusion: Multi-object tracking data generation via diffusion models. *ArXiv preprint*, abs/2312.00651, 2023.

[47] Zhili Liu, Kai Chen, Yifan Zhang, Jianhua Han, Lanqing Hong, Hang Xu, Zhenguo Li, Dit-Yan Yeung, and James Kwok. Geom-erasing: Geometry-driven removal of implicit concept in diffusion models. *ArXiv preprint*, abs/2310.05873, 2023.

[48] Yibo Wang, Ruiyuan Gao, Kai Chen, Kaiqiang Zhou, Yingjie Cai, Lanqing Hong, Zhenguo Li, Lihui Jiang, Dit-Yan Yeung, Qiang Xu, and Kai Zhang. Detdiffusion: Synergizing generative and perceptive models for enhanced data generation and perception. *ArXiv preprint*, abs/2403.13304, 2024.

[49] Kai Chen, Lanqing Hong, Hang Xu, Zhenguo Li, and Dit-Yan Yeung. Multisiam: Self-supervised multi-instance siamese representation learning for autonomous driving. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 7526–7534. IEEE, 2021.

[50] Kai Chen, Zhili Liu, Lanqing Hong, Hang Xu, Zhenguo Li, and Dit-Yan Yeung. Mixed autoencoder for self-supervised visual representation learning. In *CVPR*, 2023.

[51] Zhili Liu, Jianhua Han, Lanqing Hong, Hang Xu, Kai Chen, Chunjing Xu, and Zhenguo Li. Task-customized self-supervised pre-training with scalable dynamic routing. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 1854–1862. AAAI Press, 2022.

[52] LIU Zhili, Kai Chen, Jianhua Han, HONG Lanqing, Hang Xu, Zhenguo Li, and James Kwok. Task-customized masked autoencoder via mixture of cluster-conditional experts. In *ICLR*, 2023.

[53] Jianhua Han, Xiwen Liang, Hang Xu, Kai Chen, Lanqing Hong, Chaoqiang Ye, Wei Zhang, Zhenguo Li, Xiaodan Liang, and Chunjing Xu. Soda10m: Towards large-scale object detection benchmark for autonomous driving. *ArXiv preprint*, abs/2106.11118, 2021.

[54] Kaican Li, Kai Chen, Haoyu Wang, Lanqing Hong, Chaoqiang Ye, Jianhua Han, Yukuai Chen, Wei Zhang, Chunjing Xu, Dit-Yan Yeung, et al. Coda: A real-world road corner case dataset for object detection in autonomous driving. *ArXiv preprint*, abs/2203.07724, 2022.

Appendix

A More Analyses

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 [5, 39], 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 4, 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.

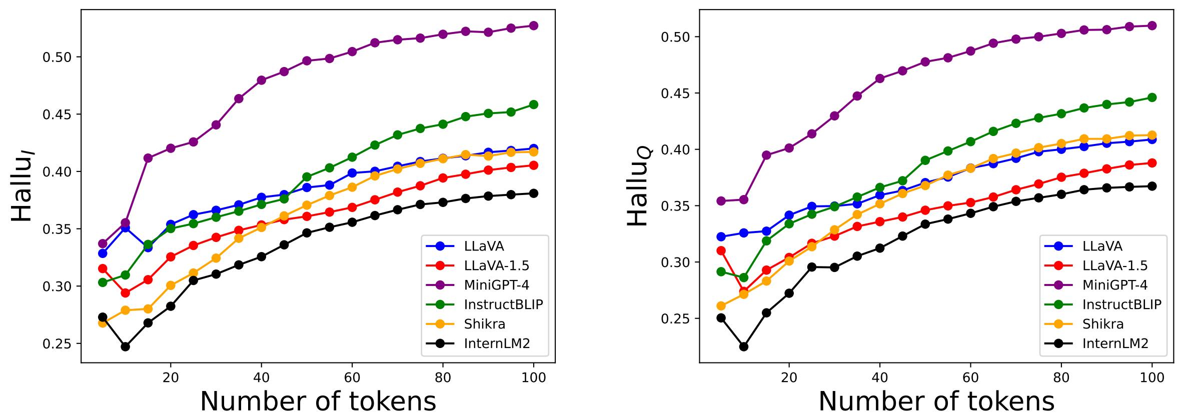


Figure 4: **Trends of the hallucination rates** of the image-level (left) and question-level (right) evaluations for different LVLMs with respect to the number of tokens in the model responses.

B More on Hallucination Mitigation

Ablation study. We conducted an ablation study on a general description (*i.e.*, "describe the image"). As shown in Table 6, the combination of both triplet description prompting and disability of visual input contributed to the effectiveness of our hallucination mitigation approach.

Mitigation	GPT-4 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
Ours (eyes-close)	46.50	49.19	54.59	53.03
Ours (triplet)	45.65	45.16	59.35	55.57
Ours (triplet+ eyes-close)	**44.14**	**42.96**	**51.19**	**47.12**

Table 6: **Ablation study on MiniGPT-4 [3].** The best results under each column are **boldfaced**.

Mitigation on LLaVA-1.5. To provide a comprehensive evaluation, we conducted additional experiments using another LVLM (*i.e.,* LLAVA-1.5) to mitigate hallucinations. The results presented in Table 7 indicate that both the general description and triplet description methods show an improved reduction in hallucinations, further highlighting the effectiveness of our mitigation approach.

Baseline Comparison. Current methods for mitigating hallucinations typically involve retraining, integrating external detection models, and devising decoding strategies. Compared to existing works, our approach is a plug-and-play method that neither requires costly retraining nor relies on external models. To make a more fair comparison, we

Mitigation	GPT-4 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	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 7: **Further experiment conduct on LLaVA-1.5 [14].** The best results under each column are **boldfaced**.

conducted experiments with LogicCheckGPT [42], a training-free approach that addresses hallucinations through prompting interactions with the help of GPT-3.5. Under the cost consideration, an evaluation was conducted only with the NLI judge. The results indicate that our method outperforms LogicCheckGPT, highlighting its effectiveness in mitigating hallucinations as shown in Table 8.

Model	Mitigation	$\text{Hallu}_\text{I} \downarrow$	$\text{Hallu}_\text{Q} \downarrow$
LLaVA-1.5	w/o Mitigation	53.84	52.06
	LogicCheckGPT	51.10	50.84
	Ours (triplet+ eyes-close)	**48.97**	**48.40**
MiniGPT-4	w/o Mitigation	55.93	59.94
	LogicCheckGPT	52.34	53.04
	Ours (triplet+ eyes-close)	**51.19**	**47.12**

Table 8: **Hallucination mitigation results on LLaVA-1.5 [14] and MiniGPT-4 [3] with baseline comparison.** The best results under each column are **boldfaced**.

Qualitative comparison with LogicCheckGPT [42] is provided in Figure 5.

C 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.

D 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.

E More on Prompts

E.1 Prompt for triplets extraction with GPT-4

The prompt for extracting triplets in the answer generated by LVLMs is illustrated in Figure 6.

E.2 Prompt for question generation with GPT-4V

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

Question: What season could it be based on the attire of the people?

MiniGPT-4: Based on the attire of the people in the image, it appears to be a warm season. The majority of the people are wearing light clothing, such as shorts and t-shirts, which are suitable for warm weather. Additionally, some of the people are wearing sandals, which are also a common footwear choice for warm weather. The architecture of the buildings in the image also supports the idea of a warm season. The buildings have large windows and open balconies, which allow for good air circulation and natural ventilation, making them suitable for warm weather. The presence of greenery, such as the plants on the balconies, also suggests a warm season, as greenery tends to thrive in warm weather. Overall, the image appears to be set in a warm season.

LogicCheckGPT [MiniGPT-4]: The image appears to be of a group of people standing on a cobblestone street. There are several tall buildings in the background with balconies and windows. The people are dressed in warm clothing, indicating that it is likely fall or winter. Some of the people are wearing scarves and jackets, while others have on lighter sweaters or jackets. The scene is lit by streetlights, and the buildings have a white, sandy-colored facade.

Ours (triplet+eyes-close) [MiniGPT-4]: The people in the image are dressed in warm clothing, indicating that it is likely a winter scene. The buildings in the background are made of stone and have large windows, suggesting that they are old and possibly historic. The people are standing on the sidewalk, which is made of cobblestones and has a drainage grate in the middle. There are no cars or other vehicles in the scene, indicating that it is a pedestrian area. The overall atmosphere of the scene is quiet and still, with the people standing in a line and not interacting with each other. Based on this information, it is likely that the scene is set in the winter season.

Question: What might the person's relationship be with the animal?

LLaVA-1.5: Based on the image, the person is a woman crouching down and holding a small kitten in her arms. It is likely that she has a close relationship with the kitten, possibly as its owner or caretaker. The woman's actions of holding the kitten and providing care suggest that she is familiar with the animal and is attentive to its needs. This could indicate that she is responsible for the kitten's well-being and is nurturing its growth and development.

LogicCheckGPT [LLaVA-1.5]: Based on the image, the woman is holding a small kitten in her arms. It is likely that she has a close relationship with the kitten, possibly as its owner. The woman's actions of holding the kitten suggest that she is familiar with the animal and is attentive to its needs. This could indicate that she is responsible for the kitten's well-being and is nurturing its growth and development.

Ours (triplet+eyes-close) [LLaVA-1.5]: The person's relationship with the animal is likely that of the owner or caretaker. The presence of the small black kitten lying on the floor and the person holding a bottle, which could be a baby bottle for the kitten, suggests that the person is providing care and nourishment to the kitten. Additionally, the presence of a bowl in the image supports the idea that the person is taking care of the kitten, as it could be used for feeding the kitten or providing water. This indicates a close and nurturing relationship between the person and the kitten.

Figure 5: **More illustrations demonstrating hallucination mitigation with comparison to LogicCheckGPT.** Hallucinated content is highlighted in Red.

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 6: **Prompt for triplets extraction with GPT-4.**

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.**

E.3 Prompt for GPT-4 Judge

The prompt for our proposed GPT-4 judge method is illustrated in Figure 8.

```
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 8: **Prompt for the GPT-4 Judge method.**

E.4 Prompt for LVLMs in Evaluation

The prompt for generating responses from LVLMs for evaluation is the combination of a question and the corresponding image.

F Human Annotation Guideline

The detailed guidelines of our human evaluation tasks are shown in Table 9. 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 9: **Detailed human evaluation instructions.**

G More Discussion

Future works. Currently, the proposed triplet-level evaluation is primarily deployed on LVLMs, whose extension to diffusion models [43, 44, 45, 46, 47, 48] is feasible, while for the hallucination mitigation proposed in §5.4 can be further enhanced by utilizing stronger vision encoder [49, 50, 51, 52] and visual tools (*e.g.*, object detectors [53, 54]) to better extract visual information for LLM reasoning.