AUTOHALLUSION: Automatic Generation of Hallucination Benchmarks for Vision-Language Models

Anonymous ACL submission

Abstract

Large vision-language models (LVLMs) hallucinate: certain context cues in an image may trigger the language module's overconfident and incorrect reasoning on abnormal or hypothetical objects. Though a few benchmarks have been developed to investigate LVLM hallucinations, they mainly rely on hand-crafted corner cases whose fail patterns may hardly generalize, and finetuning on them could undermine their validity. These motivate us to develop the first automatic benchmark generation approach, AUTOHALLUSION, that harnesses a few principal strategies to create diverse hallucination examples. It probes the language modules in LVLMs for context cues and uses them to synthesize images by: (1) adding objects abnormal to the context cues; (2) for two co-occurring objects, keeping one and excluding the other; or (3) removing objects closely tied to the context cues. It then generates image-based questions whose ground-truth answers contradict the language module's prior. A model has to overcome contextual biases and distractions to reach correct answers, while incorrect or inconsistent answers indicate hallucinations. AUTOHALLUSION enables us to create new benchmarks at the minimum cost and thus overcomes the fragility of hand-crafted benchmarks. It also reveals common failure patterns and reasons, providing key insights to detect, avoid, or control hallucinations. Comprehensive evaluations of top-tier LVLMs, e.g., GPT-4V(ision), Gemini Pro Vision, Claude 3, and LLaVA-1.5, show a 97.7% and 98.7% success rate of hallucination induction on synthetic and real-world datasets of AUTOHALLUSION, paving the way for a long battle against hallucinations.

1 Introduction

Large vision-language models (LVLMs) (Openai, 2023; Liu et al., 2023c) bring powerful tools for content generation (Lian et al., 2024), autonomous

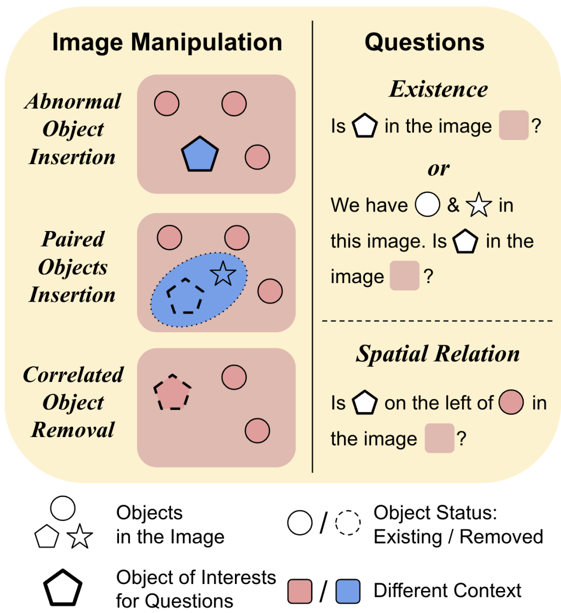


Figure 1: **AUTOHALLUSION:** We propose three image manipulation strategies to induce hallucinations: *abnormal object insertion*, *paired object insertion*, and *correlated object removal*, which trigger the conflicts between the images and LVLM priors. Given generated images, we ask LVLMs questions on object existence and their spatial relations for visual question answering.

driving (Chen et al., 2024), and robotics (Brohan et al., 2023; Guan et al., 2024b). However, hallucinations (Zhang et al., 2023), i.e., LVLM-generated responses contain information not present in the visual content, raise an alarm and limit LVLMs' applications. Hallucinations occur when LVLMs' perception and reasoning over-rely on the strong priors of their language modules while ignoring the visual sensory inputs (Guan et al., 2024a).

It is critical for the research community to collect these cases and investigate the reasons behind them. With sufficient hallucination examples, finetuning LVLMs on them with the original training data may reduce hallucinations and alleviate those biases. However, crafting those cases in previous work requires expensive human labor and is highly time-consuming (Jiang et al., 2024; Rohrbach et al., 2018; Li et al., 2023b; Han et al., 2024; Guan et al., 2024a). Moreover, it is unclear whether those handcrafted examples are rare corner cases or indicate general fail patterns. Without an in-depth understanding of the common mechanism generating them, it is hard to extract useful insights to improve LVLMs. On the other hand, finetuning on those small benchmarks without sufficient representative examples may lead to overfitting.

To address those challenges, we develop an automated pipeline, AUTOHALLUSION, to generate diverse hallucination cases and mass-produce them at the minimum cost of human efforts. To generate (image, question) pairs that can trigger the hallucinations of LVLMs, we take a reverse-engineering path: It starts from exploring output answers due to hallucinations, by probing LVLMs' language modules to allocate the strong language priors on certain objects or their contextual relations. It then creates (1) an image containing objects that contradict the probed priors (the presumed answers), and (2) questions on two types of conflicts, the existence of contextual-related objects and their spatial relationships. If the LVLM reasoning is biased or dominated by the language prior, it tends to generate incorrect or inconsistent responses conflicting with the ground truth in the images, hence the hallucinations. We provide an optimization formulation and develop three principal strategies, *abnormal object insertion*, *paired object insertion*, and *correlated object removal*, to manipulate the objects in a scene and thus create images conflicting with the language prior, as illustrated in Figure 1.

The detailed designs of these hallucination strategies are inspired by *schema* (DiMaggio, 1997; Boutyline and Soter, 2021; Rumelhart, 2017) from cognitive science. *Schema* refers to the tendency of humans to organize information and interpret the world based on patterns of past experiences[1]. Following its concept, *irregular schema* with *cognitive dissonance* (Aronson, 1969; Harmon-Jones and Mills, 2019), e.g., an octopus in front of a monitor, and *breaking a schema* with *expectancy violation* (Burgoon, 1993; Burgoon and Hale, 1988), e.g., the absence of a keyboard and a mouse in front of a monitor, can both induce contradictions and discomforts in the memory. The three strategies reveal common patterns and mechanisms of how hallucinations are generated, hence providing critical insights to detect, combat, avoid, or control hallucinations of LVLMs.

[1]For example, it is much more common to see a keyboard and a mouse in front of a monitor rather than an octopus.

Main contributions: Inspired by an analogy to human cognition in terms of *schema*, we investigate the mechanism of hallucinations in LVLMs by reverse-engineering (image, question) pairs with probed language priors and biases. We develop AUTOHALLUSION that synthesizes images by manipulating the objects in the scenes to conflict with LVLMs' memory (i.e., its language priors), and generates questions about the conflicts. The novelties of our work can be summarized as:

- We propose the first automatic generation approach of hallucination benchmarks, with a high-level formulation and three principal strategies, inspired by *schema* in cognitive science, to trigger LVLM hallucinations.
- We develop novel probing methods to extract and investigate the contextual biases in the language priors that cause hallucinations. We further introduce two evaluation metrics to detect hallucinations.
- We evaluate SOTA LVLMs, including GPT-4V(ision), Gemini Pro Vision, Claude 3, and LLaVA-1.5, on benchmarks by AUTOHALLUSION. It achieves success rates of 97.7% and 98.7% of inducing LVLM hallucinations on synthetic and real-world data.

2 Related Work

Vision-Language Models (VLMs). The recent increase in large language models (LLMs), including GPT-3 (Brown et al., 2020), PaLM (Chowdhery et al., 2023), and BLOOM (Le Scao et al., 2023), has significantly improved natural language processing. LLaMA (Touvron et al., 2023) further advanced this field, and models like Alpaca (Taori et al., 2023), inspired by InstructGPT (Ouyang et al., 2022) and ChatGPT, utilized human-annotated data to refine LLaMA, enhancing its interaction abilities. Additionally, Large Visual Language Models (LVLMs) such as GPT-4 (Achiam et al., 2023), Flamingo (Alayrac et al., 2022), Bard (AI, 2023), MiniGPT-4 (Zhu et al., 2023), InstructBLIP (Dai et al., 2024), and LLaVA (Liu et al., 2024b) have developed. These models combine visual and language processing to manage both textual and visual inputs and produce textual outputs. Their architecture generally includes a visual encoder (often based on CLIP (Radford et al., 2021)), a modality connection module, and an LLM. LVLMs excel in generating text descriptions from images and multi-modal learning through pre-training on image-text pairs (Liu et al., 2024a) and instruction-tuning with various tasks (Liu et al., 2023a; Ouyang et al., 2022; Xu et al., 2023). However, addressing hallucinations in their textual outputs remains a challenge, emphasizing the need for reliability and accuracy in real-world applications.

Benchmarks. Several benchmarks have been developed to assess hallucination in VLMs in various aspects. CHAIR (Rohrbach et al., 2018) evaluates object hallucination by measuring word accuracy against ground-truth sentences and segmentation for 80 MSCOCO objects. POPE (Li et al., 2023b) improves upon CHAIR for better stability and flexibility while OpenCHAIR (Ben-Kish et al., 2023) extends CHAIR to open-vocabulary settings. HallusionBench (Guan et al., 2024a) targets visual commonsense and reasoning with 455 visual-question control pairs. Hal-Eval (Jiang et al., 2024) introduces and focuses on event hallucination while CorrelationQA (Han et al., 2024) examines the impact of spurious visual inputs. Our work differs from previous benchmarks by using an auto-generated hallucination approach, synthesizing visual hallucination cases through contextual influences.

Object Hallucination. Large Vision Language Models (LVLMs) hold great potential but struggle with object hallucination, generating incorrect descriptions that include nonexistent objects or omit key details. This problem can adversely affect applications in robotics (Wu et al., 2024; Liu et al., 2023b), medical imaging (Wang et al., 2023; Hu et al., 2023), and human-computer interaction (Brie et al., 2023). Object hallucination in LVLMs manifests as fictional objects, false attributes, or inaccurate relationships between objects (Gunjal et al., 2023; Zhai et al., 2023). Previous methods, like fine-tuning smaller multimodal models (Biten et al., 2022; Kim et al., 2023), are less effective for LVLMs due to their distinct architectures. Recent efforts focus on improving dataset quality for fine-tuning (Li et al., 2023a; Liu et al., 2023a), but acquiring such data remains labor-intensive. Metrics like CHAIR (Rohrbach et al., 2018) and POPE (Li et al., 2023b), which assess caption relevance and hallucination levels, are crucial for evaluation. Standard text quality metrics can be misleading, as high scores may still correlate with significant hallucination. In this paper, we investigates contextual biases in language priors causing hallucinations and introduces two new metrics for more effective detection.

3 Problem Formulation

Pronounced bias in LLMs hinders the reasoning capability of LVLMs, resulting in hallucinations from the images (Guan et al., 2024a). Inspired by this, we target the biases in LLMs to induce hallucinations in LVLMs.

Definitions and Objective. Our objective is to find things that are correlated in the LLM but not present in the picture to induce hallucinations in LVLMs. Let $f_{\texttt{LVLM}}$, $f_{\texttt{LLM}}$ denote the LVLM and its LLM component, respectively. $f_{\texttt{LVLM}}(\textit{image}, \textit{query})$ can take an image-query pair as inputs, and $f_{\texttt{LLM}}(\textit{context}, \textit{query})$ can take a text-only context-query pair as inputs. We use sets as universal representations for the images and texts and detailed as below.

We denote $\mathcal{V}$ as the set of all contextual elements in an image I, where each element can be an object, an attribute associated with an object, or the relation between/among multiple objects, etc.[2] These elements in the set can be considered as a statement, which could be either affirmative or negative. Similarly, for text modality, we denote $\mathcal{Q}$ as the set containing objects of interest for questions and $\mathcal{C}$ as the set of objects in this scene for context. We use a mapping function $T(\cdot)$ to transform a set of contextual elements into a text, which can be either a description from $\mathcal{C}$ or a query question from $\mathcal{Q}$.

Finally, we introduce the contextual distance $d[\cdot, \cdot]$ between two descriptions or texts. When two pieces of text convey similar information or affirm each other, the contextual distance d is considered small; otherwise, the contextual distance is large. Let $y_{\mathcal{Q}}$ be the ground truth answer set with respect to the query set $\mathcal{Q}$ given the image I. The objective function can be formulated as follows:

$$\max_{I,\mathcal{Q},\mathcal{C}} \quad d[f_{\texttt{LVLM}}(I, T(\mathcal{Q})), y_{\mathcal{Q}}] \tag{1}$$

$$\begin{aligned} s.t. \quad & d[f_{\texttt{LVLM}}(I, T(\mathcal{Q})), f_{\texttt{LLM}}(T(\mathcal{C}), T(\mathcal{Q}))] \leq \epsilon, \\ & \mathcal{C} \subseteq \mathcal{V}, \mathcal{Q} \cap \mathcal{C} = \emptyset. \end{aligned} \tag{2}$$

The objective function (1) maximizes the distance between the generated text $f_{\texttt{LVLM}}$ and $y_{\mathcal{Q}}$ to produce hallucination. To leverage and probe the

[2]similar to the visual genome dataset.

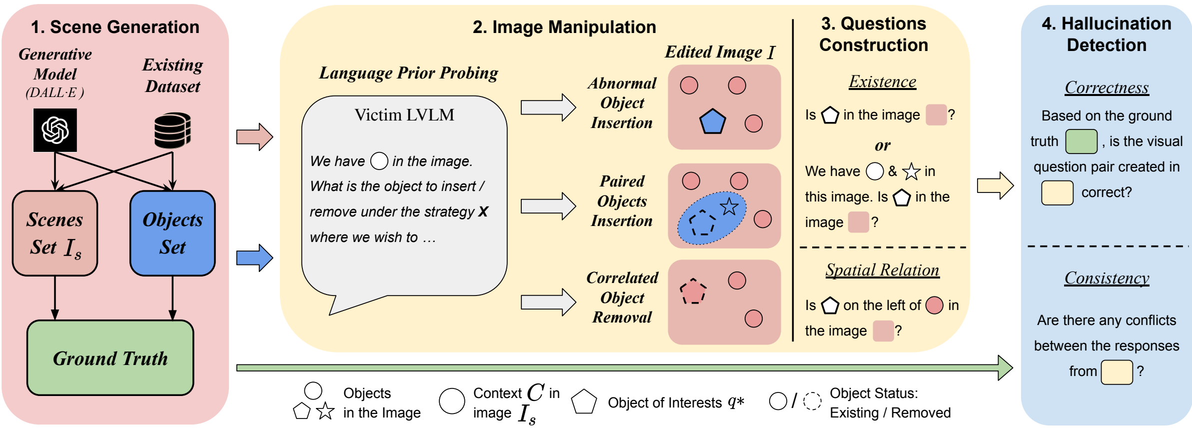


Figure 2: **Overview of AutoHallusion.** We first automatically generate the scenes set and objects set (pink). After that, we use text to probe the language prior of the victim LVLM and then propose three manipulation strategies to induce hallucination in scene images(yellow). We finally use two metrics to detect hallucinations (blue).

bias in the language components f_{LLM} of the victim LVLM, we use constrain (2) to control the discrepancies between responses, with and without visual input, within a tolerance ϵ. This ensures that the answer is dominated by the prior language component rather than the visual component. The visual information $\mathcal{V}$ from the image I provided to f_{LVLM} is partially converted to a text, $T(\mathcal{C})$, as the input to f_{LLM}, therefore $\mathcal{C} \subseteq \mathcal{V}$.

Remark. It is important for the language component f_{LLM} to have the constraint $\mathcal{Q} \cap \mathcal{C} = \emptyset$. If the interested element from $\mathcal{Q}$ is directly given in the context $\mathcal{C}$, it would be difficult to exploit the bias of f_{LLM} since it is mostly likely to answer based on the provided context $\mathcal{C}$. For example, if a fact is directly given in the prompt, it is hard for the model to make a contradictory claim. In addition, $\mathcal{Q}$ is not required to be the subset of $\mathcal{V}$ since we can ask questions on objects that are not included in the image I.

Approach. It is hard to optimize I, $\mathcal{Q}$, and $\mathcal{C}$ by directly optimizing Eq. (1). Instead, we probe the LVLM and the language prior from its LLM component to determine $(\mathcal{Q}, \mathcal{C})$ such that the elements in $\mathcal{Q}$ are highly likely (or unlikely, depending on the attack strategy) to be present with $\mathcal{C}$ in the same scene. Such bias in the language prior helps us achieve the constraint (2). This ensures that the language prior is strong and highly confident on the co-occurrence of $(\mathcal{Q}, \mathcal{C})$, i.e., $\Pr(\mathcal{Q} \mid \mathcal{C}) \leq \delta$ ($\mathcal{Q}$ is abnormal given $\mathcal{C}$) or $\Pr(\mathcal{Q} \mid \mathcal{C}) \geq 1 - \delta$ ($\mathcal{Q}$ is hypothetical given $\mathcal{C}$), where $\Pr(\mathcal{Q} \mid \mathcal{C})$ is the probability of the existence of elements in $\mathcal{Q}$ given the presence of $\mathcal{C}$ and δ is the confidence level. If the assumption on $(\mathcal{Q}, \mathcal{C})$ pairs that the LVLM reasoning is dominated by its language prior, *i.e.* Eq. (1) holds true, we can create I from such $(\mathcal{Q}, \mathcal{C})$ pairs to maximize the discrepancy in Eq. (1).

4 Methodology

The overall pipeline of our methodology is presented in Fig. 2. We break down the automated procedure of creating hallucination cases into 4 stages: scene generation, image manipulation, question construction, and hallucination detection. Questions constructed to induce potential hallucination cases vary depending on these strategies, mainly focusing on object existence and spatial relations between the target object and others. We detect hallucinations through correctness and consistency among answers generated by the victim LVLM.

4.1 Scene Generation

First, we want to create a scene image I_s with a strong context $\mathcal{C}$ so that it would be easier to extract bias and incur hallucination. Given a random scene name or a brief description, we use the target LVLM to generate and expand on the contextual elements $\mathcal{C}$ within the scene. With these descriptions and details, we employ a diffusion model or an image generation model like DALL-E-3 (OpenAI, 2023) to create an image I_s rich in context, incorporating the provided scene information and relevant objects that are listed in the context $\mathcal{C}$. Alternatively, I_s can be obtained from an existing dataset, assuming the images are coherent, natural, and contain several correlated elements. We use Owl-ViT (Minderer et al., 2022) to ground the contextual elements of I_s and verify the context $\mathcal{C}$.

4.2 Image Manipulation

Once we have a scene image I_s rich in context, we want to use $\mathcal{C}$ to probe the LLM component $f_{\texttt{LLM}}$ of the victim model and find a target object, which is used to modify I_s. This target object is not only used to manipulate I_s, but also used to construct the questions $\mathcal{Q}$. Once we find a suitable $\mathcal{Q}$ based on $\mathcal{C}$, we can modify I_s and manipulate the target object to obtain the final I.

Our hallucination attack focuses one contextual element q^* retrieved from the query set $\mathcal{Q}$. Since $\mathcal{Q}$ is not bounded to all the visual elements $\mathcal{V}$ from the image, the modification can be either object insertion or removal. Our manipulation strategies are explained as follows:

4.2.1 Abnormal Object Insertion

The abnormal object insertion strategy tries to insert an object not related to the existing contextual elements into the scene image I_s. For example, given an image of an office scene, a suitable abnormal object that contradicts this context could be a cooking pot.

The query question q^*, which is also the abnormal object for insertion, should have the **maximum sum of distances** between its language prior and the ground truth information across all contextual elements in $\mathcal{C}$. We bound the retrieval process as:

$$q^* = \arg\max_{q \in \mathcal{Q}} \sum_{c \in \mathcal{C}} d(f_{\texttt{LLM}}(T(c), T(q)), y_q). \quad (3)$$

In practice, we use DALL-E 3 (OpenAI, 2023) to create this image for the abnormal object or choose the object's image from an existing database. To guarantee the insertion is successful and does not introduce any artifacts, we simply stitch the object into the scene image after removing the background of the object image (Nader, 2021) instead of using diffusion or an in-painting method.

4.2.2 Paired Object Insertion

The paired object insertion strategy uses target LVLM to determine the paired objects with a strong correlation, like coffee makers and coffee beans. In this strategy, we insert only one of two objects from the pair and ask questions about the other.

We formulate this image manipulation process into finding the query question q^* with the **minimum element-wise distance** between its language prior and the ground truth information among all contextual elements in $\mathcal{C}$:

$$q^* = \arg\min_{q \in \mathcal{Q}} \min_{c \in \mathcal{C}} d(f_{\texttt{LLM}}(T(c), T(q)), y_q) \quad (4)$$

4.2.3 Correlated Object Removal

The correlated object removal strategy removes the existing object from the generated scene image I_s, while the removed object has a strong correlation with multiple contextual elements within I_s. We query such an adversary object q^* by searching for the object with the **minimum sum of distances** between its language prior and the ground truth information across all contextual elements in $\mathcal{C}$:

$$q^* = \arg\min_{q \in \mathcal{Q}} \sum_{c \in \mathcal{C}} d(f_{\texttt{LLM}}(T(c), T(q)), y_q) \quad (5)$$

Intuition. The purpose of two types of insertion is to add an abnormal element that the model will ignore given the strong context and to insert one of a correlated object pair so the model will hallucinate about the other, respectively. For the removal strategy, we aim to identify and erase the object that has the strongest correlation with the context or the shortest sum of contextual distances to the ground truth of the scene given the query.

4.3 Question Construction

We mainly consider two types of questions: the existence of the object and the spatial relation between the objects.

In the existence questions, we ask whether the target object q^* is present in the image. These questions are repeated multiple times, with varying levels of details mentioned in the prompt. For example, we ask the same victim model to generate an image caption and add this text in front of the query question because such supplementary information may serve as another source of language prior that misleads the victim model. In addition, we also ask existence questions on objects that are missing in the image caption generated by the victim model because it has a higher probability of missing this object again in the existence question.

In the spatial relation question, we ask about the relative positions of the target object and the scene objects. Given the bounding boxes, it's easy to obtain the following spatial relations: *Left*, *Right*, *Above*, *Below*, *Front* (when the perturbed object overlaps with the scene object). Spatial relation questions are asked with multiple levels of contextual information from the image, including vanilla (no extra information), single and concatenated object-level description, and the detailed caption for the whole image, all of which are generated by the victim model.

4.4 Hallucination Detection

We use GPT-4V-Turbo (Openai, 2023) to evaluate the correctness of the predicted answer by the victim model and the ground truth. There are two criteria to determine whether hallucination occurs with different levels of reliability:

1. **Correctness:** Since we know the ground truth existence and relations of objects, we can easily determine the correctness of the visual question pairs. This criterion is the most straightforward, but it does not account for any generation errors or background-removal artifacts from the pipeline. If some of the steps fail, the ground truth may not be reliable.
2. **Consistency:** In this criterion, we want to consider the consistency of the model outputs, which does not rely on whether the ground truth is accurate. For example, if we ask about the existence of an object and get different responses, we are certain that one of the responses is hallucinating. We divide the inconsistency hallucination into two categories: **(1) Response Conflict** happens when LVLMs fail to give consistent answers to questions with different levels of supplementary information provided, and **(2) Local-Global Conflict** occurs when LVLMs fail to provide answers about the object of interest (local) that are consistent with the caption describing the image related to that object.

5 Evaluation and Metrics

5.1 Implementation Details

Data Preparation. To obtain all the scene images and object images for insertion, we either generated those images with image generation models like DALL-E-3 (OpenAI, 2023), or use existing datasets. For image generation, we first use LVLM to fill in more details of the scene with objects for better generation results. For real-world data, we use the validation dataset from the Common Objects in Context (COCO) dataset (Lin et al., 2014). We randomly select 126 samples with sufficient contextual elements provided in the image and around $5,000$ object images segmented from raw images. We edit the scene image by inserting objects retrieved from the database, thinking

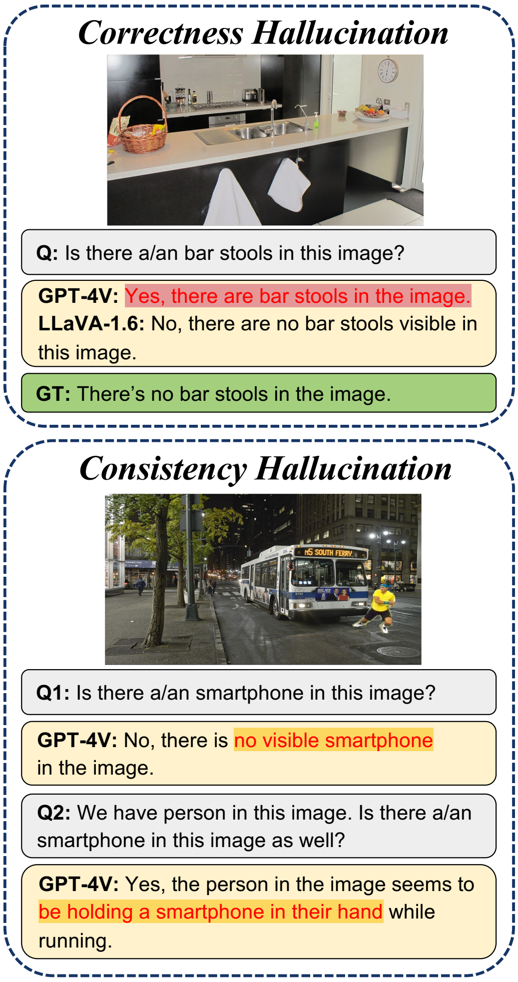


Figure 3: **Hallucination Cases created by AUTOHALLUSION.** We highlight hallucination context made by *Correctness* and *Inconsistency*

about the correlated object for the given object, or removing them from the scene.

In Appendix D, we provide showcases for both data preparation methods, including the initial scene and object images and those images after manipulation, which are either generated by DALL-E-2 (OpenAI, 2023) or queried from the real-world image dataset.

Victim LVLMs. We conduct extensive experiments using the proposed AUTOHALLUSION across the following models: GPT-4V-Turbo (Yang et al., 2023), LLaVA-1.5 (Liu et al., 2023c), Claude 3 (Team, 2024), Gemini Pro Vision (Team, 2023), and miniGPT4 (Zhu et al., 2023).

Implementation Details. We generate 200 cases for each experiment. By default, all scene and edited images are 1024×1024 and inserted objects are 200×200 for synthetic data. For real-world dataset, we loop over all scene images with proper resizing to fit the input of image models.

Manipulation Strategies	LVLMs	Synthetic Data					Real-World Data				
		Overall ASR	Overall MASR	Overall CASR	Exi. ASR	Sp. ASR	Overall ASR	Overall MASR	Overall CASR	Exi. ASR	Sp. ASR
Abnormal Obj. Insertion	GPT-4V-Turbo (Yang et al., 2023)	96.0	80.0	92.5	93.0	78.1	100.0	98.4	98.4	97.6	97.5
	Gemini Pro Vision (Team, 2023)	97.0	90.5	90.0	84.5	89.1	100.0	100.0	97.6	97.6	94.3
	Claude Pro Vision (Team, 2024)	97.4	90.7	96.0	95.3	92.3	100.0	100.0	100.0	100.0	98.4
	LLaVA-1.5 (Liu et al., 2023c)	97.7	94.2	94.0	97.9	96.2	100.0	100.0	98.9	98.6	95.9
	miniGPT4 (Zhu et al., 2023)	98.1	95.1	98.0	98.1	97.1	100.0	100.0	97.9	98.0	96.1
Paired Obj. Insertion	GPT-4V-Turbo (Yang et al., 2023)	99.5	93.5	97.0	91.5	81.7	100.0	100.0	99.2	99.2	100.0
	Gemini Pro Vision (Team, 2023)	100.0	100.0	99.5	99.5	85.7	100.0	99.2	100.0	99.2	90.4
	Claude 3 (Team, 2024)	100.0	99.0	99.0	99.0	95.5	100.0	99.2	100.0	97.6	99.2
	LLaVA-1.5 (Liu et al., 2023c)	99.7	95.1	98.9	97.6	81.8	99.7	98.5	99.3	94.5	97.8
	miniGPT4 (Zhu et al., 2023)	100.0	99.8	100.0	99.1	83.9	100.0	100.0	99.5	99.5	99.8
Correlated Obj. Removal	GPT-4V Turbo (Yang et al., 2023)	93.0	84.0	84.0	69.5	85.5	94.4	88.0	84.0	75.2	85.4
	Gemini Pro Vision(Team, 2023)	95.0	92.0	93.0	77.0	91.1	96.8	95.2	92.0	77.6	94.2
	Claude 3(Team, 2024)	99.0	98.0	89.0	92.0	88.5	98.4	98.4	94.4	96.0	89.6
	LLaVA-1.5(Liu et al., 2023c)	97.1	88.9	87.4	70.8	87.4	93.1	97.6	94.6	78.1	95.7
	miniGPT4 (Zhu et al., 2023)	96.7	90.1	91.5	72.9	86.7	97.8	96.3	89.1	76.9	87.8

Table 1: **Evaluation results of SOTA LVLMs with our AUTOHALLUSION on synthetic and real-world data.** Our proposed three manipulation strategies achieved high success rates (the higher the better) on synthetic and real-world data.

For a given generated scene image I_s, we use the object detection model (Minderer et al., 2022) to detect and segment all candidate contextual elements for removal from the image. We use the generative image model DALL-E-2 (Ramesh et al., 2022) to in-paint the chosen object for removal.

5.2 Evaluation Metrics

Apart from the overall **Attack Success Rate (ASR)** of each evaluation category, we mainly use the following evaluation metrics to determine whether hallucination generation is successful:

Manipulation Attack Success Rate (MASR): We compare the generated response with the ground truth generated based on the intention of the image generation and editing. However, it is possible that the ground truth of the image is not accurate due to failure during image generation and editing.

Conflict Attack Success Rate (CASR): We ask a set of questions and try to find conflicts among all responses to those visual questions. Such inconsistency will guarantee that one of the conflicting responses must have been hallucinated and provided an incorrect answer.

5.3 Main Results

Table 1 summarizes the performance of victim LVLMs under our three attack strategies using synthetic and real-world datasets. We achieve high ARS with all three proposed attack strategies in both datasets, showing the effectiveness of our approach to induce hallucinations.

We have the following key observations: 1) Strategies probing inserted objects (Abnormal Object and Paired Object Insertion), achieve higher hallucination attack success rates than those probing absent objects (Correlated Object Removal strategy); 2) Questions probing the existence of objects are more effective to cause hallucinations than questions probing spatial relations; 3) GPT-4V-Turbo is the most robust to hallucination attacks among all victim LVLMs; 4) Our method achieved even higher attack success rates across all LVLMs in the real-world dataset than synthetic data. We hypothesize this comes from LVLMs lack of ability to address the complexity and diversity within the real-world data, which causes its higher vulnerability to our attack strategies when using real-world data. For more experimental results, please refer to Appendix A.

5.4 Ablation Studies

Object Sizes. Table 2 shows results for different object sizes from 100×100 to 400×400 using an abnormal object insertion strategy with GPT-4V-Turbo, while AUTOHALLUSION generally uses 200×200. The findings indicate that larger objects reduce hallucinations, including those from image manipulation and response conflicts. Similar patterns are evident in questions probing existence and spatial relationships. LVLMs are more vulnerable to smaller perturbed objects, as they struggle to encode small images into tokens. However, we attribute this phenomenon comes from visual illusions made by the failure of visual encoders of LVLMs, instead of hallucinations targeting the rea-

Obj. Size	Overall			Existence			Spatial Relation		
	Overall ASR	Overall MASR	Overall CASR	Exi. ASR	Exi. MASR	Exi. CASR	Sp. ASR	Sp. MASR	Sp. CASR
100 × 100	98.0	90.0	97.5	97.0	78.5	96.0	87.5	80.6	70.0
200 × 200	96.0	80.0	92.5	93.0	62.0	88.5	78.1	71.2	60.6
300 × 300	93.5	75.0	85.5	87.0	54.0	80.5	76.3	69.4	45.0
400 × 400	89.5	68.5	79.0	81.0	43.5	74.0	65.6	53.8	41.9

Table 2: **Ablation on the size of the objects with abnormal object insertion using GPT-4V-Turbo.**

Alignment	Overall			Existence			Spatial Relation		
	Overall ASR	Overall MASR	Overall CASR	Exi. ASR	Exi. MASR	Exi. CASR	Sp. ASR	Sp. MASR	Sp. CASR
Abnormal	96.0	80.0	92.5	93.0	62.0	88.5	78.1	71.2	60.6
Random	98.5	82.0	93.5	91.5	50.5	89.0	84.0	74.9	59.4
Same	93.0	65.5	90.0	88.0	27.5	85.5	83.1	70.9	62.2

Table 3: **Ablation on object-scene alignments with abnormal object insertion using GPT-4V-Turbo.**

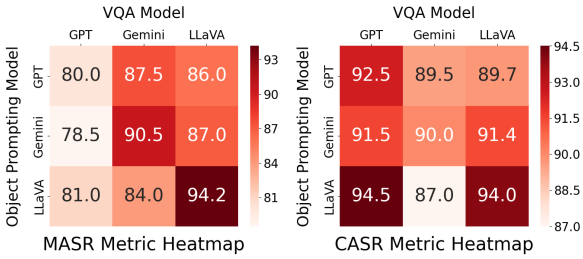


Figure 4: **Ablation on using different LVLMs for object prompting and VQA tasks.**

soning abilities of LVLMs. We selected the current object size to balance hallucination attack performance with the reduction of visual illusions.

Object Prompting and VQA Alignment. As we mentioned in Section 4.2 and 4.3, we use the same victim model to prompt objects for image manipulation and perform VQA tasks with constructed questions, which may introduce inherited biases. We conduct ablation experiments to debias and evaluate models' performance on each sub-task separately by swapping models for object prompting and VQA with abnormal object retrieval strategy. Fig. 4 shows the results using different models among GPT-4V-Turbo, Gemini Pro Vision, and LLaVA-1.5 performing abnormal object prompting and VQA tasks. Results show that models have varied performance over different metrics, like GPT-4V-Turbo is more robust to correctness hallucinations and Gemini is more robust to consistent hallucinations. Our results affirm the effectiveness of our pipeline in crafting hallucination cases with a high attack success rate, while using the same model for object prompting and VQA tasks usually causes more hallucinations due to inherited biases. We attribute this phenomenon to the diversity of the prior across different LVLMs as the VQA model may find the object prompted by other LVLMs less abnormal and it is less likely to suffer from hallucinations by this prompted object.

Object-scene Alignment. Table 3 presents results using different object retrieval policies under object insertion experiments using GPT-4V-Turbo, including abnormal (intentionally chooses irrelevant objects), random (randomly chooses objects), and same (chooses objects aligned with the existing contexts in the image). Results show that the abnormal object insertion strategy shows a significantly high ASR over questions probing the existence of perturbed objects, and the same object insertion strategy shows a greatly lower overall MASR. As the object retrieval and insertion strategy mainly affects the LVLMs' ability to identify the perturbed objects from the image, abnormal object insertion more easily causes the cognitive disorder of LVLMs, reflected by the high MASR values. On the other hand, LVLMs are more likely to make correct predictions when the perturbed objects are contextually aligned with the image, leading to a lower MASR value.

6 Conclusion

In this paper, we introduce AUTOHALLUSION, the first automatic benchmark generation approach to create diverse hallucination examples. Inspired by *schema* in cognitive science, we analyze the mechanism of how and when LVLM hallucinations are triggered. We then reverse-engineer the hallucinating images based on probed LVLMs' language priors by three principal strategies, abnormal object insertion, paired object insertion, and correlated object removal, that manipulate scene images using object insertion or removal to create conflicts with the priors. We construct textual probing methods to construct and detect hallucinations created. AUTOHALLUSION achieves a significant success rate of inducing LVLM hallucinations on manipulating both synthetic and real-world data. We will keep improving the quality of the synthesized images by inpainting techniques based on more recent text-to-image models. Meanwhile, we will explore better textual probing methods extracting more diverse contextual information within the image. We will also further investigate the causes of multi-modal hallucinations and build a more rigorous mathematical model for them.

7 Limitation

A limitation of our current image manipulation strategies lies on the object insertion, where we are using a primitive image stitch pipeline to insert prompted objects into the scene image. Though the success of this strategy is supported by the experimental results, the edited images have strong perceivable hand-crafting evidences which lower the quality of the resulted hallucinating images. Another limitation comes from the diversity of questions, as they mainly focus on objects' existence and spatial relations but have not explore the objects' attributes, e.g., color, pattern, and conditions, on which hallucinations might also emerge. We will take efforts to overcome them in our future update of AUTOHALLUSION.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Google AI. 2023. Bard: Google ai's conversational ai. https://ai.google/. Accessed: 2024-05-19.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems, 35:23716–23736.

Elliot Aronson. 1969. The theory of cognitive dissonance: A current perspective. In Advances in experimental social psychology, volume 4, pages 1–34. Elsevier.

Assaf Ben-Kish, Moran Yanuka, Morris Alper, Raja Giryes, and Hadar Averbuch-Elor. 2023. Mocha: Multi-objective reinforcement mitigating caption hallucinations. arXiv preprint arXiv:2312.03631.

Ali Furkan Biten, Lluís Gómez, and Dimosthenis Karatzas. 2022. Let there be a clock on the beach: Reducing object hallucination in image captioning. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 1381–1390.

Andrei Boutyline and Laura K Soter. 2021. Cultural schemas: What they are, how to find them, and what to do once you've caught one. American Sociological Review, 86(4):728–758.

Paul Brie, Nicolas Burny, Arthur Sluÿters, and Jean Vanderdonckt. 2023. Evaluating a large language model on searching for gui layouts. Proceedings of the ACM on Human-Computer Interaction, 7(EICS):1–37.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. arXiv preprint arXiv:2307.15818.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901.

Judee K Burgoon. 1993. Interpersonal expectations, expectancy violations, and emotional communication. Journal of language and social psychology, 12(1-2):30–48.

Judee K Burgoon and Jerold L Hale. 1988. Nonverbal expectancy violations: Model elaboration and application to immediacy behaviors. Communications Monographs, 55(1):58–79.

Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. 2024. Driving with llms: Fusing object-level vector modality for explainable autonomous driving. IEEE International Conference on Robotics and Automation (ICRA).

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. Journal of Machine Learning Research, 24(240):1–113.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in Neural Information Processing Systems, 36.

Paul DiMaggio. 1997. Culture and cognition. Annual review of sociology, 23(1):263–287.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024a. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. Preprint, arXiv:2310.14566.

Tianrui Guan, Yurou Yang, Harry Cheng, Muyuan Lin, Richard Kim, Rajasimman Madhivanan, Arnie Sen, and Dinesh Manocha. 2024b. Loc-zson: Language-driven object-centric zero-shot object retrieval and navigation. Preprint, arXiv:2405.05363.

Anish Gunjal, Jihan Yin, and Erhan Bas. 2023. Detecting and preventing hallucinations in large vision language models. In AAAI Conference on Artificial Intelligence.

Tianyang Han, Qing Lian, Rui Pan, Renjie Pi, Jipeng Zhang, Shizhe Diao, Yong Lin, and Tong Zhang. 2024. The instinctive bias: Spurious images lead to hallucination in mllms. arXiv preprint arXiv:2402.03757.

Eddie Harmon-Jones and Judson Mills. 2019. An introduction to cognitive dissonance theory and an overview of current perspectives on the theory.

Mingzhe Hu, Shaoyan Pan, Yuheng Li, and Xiaofeng Yang. 2023. Advancing medical imaging with language models: A journey from n-grams to chatgpt. arXiv preprint arXiv:2304.04920.

Chaoya Jiang, Wei Ye, Mengfan Dong, Hongrui Jia, Haiyang Xu, Ming Yan, Ji Zhang, and Shikun Zhang. 2024. Hal-eval: A universal and fine-grained hallucination evaluation framework for large vision language models. arXiv preprint arXiv:2402.15721.

Jae Myung Kim, A Koepke, Cordelia Schmid, and Zeynep Akata. 2023. Exposing and mitigating spurious correlations for cross-modal retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2584–2594.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, et al. 2023a. A large-scale dataset towards multi-modal multilingual instruction tuning. arXiv preprint arXiv:2306.04387.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. arXiv preprint arXiv:2305.10355.

Long Lian, Baifeng Shi, Adam Yala, Trevor Darrell, and Boyi Li. 2024. Llm-grounded video diffusion models. International Conference on Learning Representations (ICLR).

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

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023a. Aligning large multi-modal model with robust instruction tuning. arXiv preprint arXiv:2306.14565.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024a. A survey on hallucination in large vision-language models. arXiv preprint arXiv:2402.00253.

Haokun Liu, Yaonan Zhu, Kenji Kato, Izumi Kondo, Tadayoshi Aoyama, and Yasuhisa Hasegawa. 2023b. Llm-based human-robot collaboration framework for manipulation tasks. arXiv preprint arXiv:2308.14972.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023c. Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. Advances in neural information processing systems, 36.

M Minderer, A Gritsenko, A Stone, M Neumann, D Weissenborn, A Dosovitskiy, A Mahendran, A Arnab, M Dehghani, Z Shen, et al. 2022. Simple open-vocabulary object detection with vision transformers. arxiv 2022. arXiv preprint arXiv:2205.06230, 2.

Johnathan Nader. 2021. Background remover: Remove background from images and video using ai. https://github.com/nadermx/backgroundremover.

OpenAI. 2023. Dall·e 3: Creating images from text. https://www.openai.com/research/dall-e-3.

Openai. 2023. Gpt-4v(ision) system card.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. Advances in neural information processing systems, 35:27730–27744.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PMLR.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 1(2):3.

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

David E Rumelhart. 2017. Schemata: The building blocks of cognition. In Theoretical issues in reading comprehension, pages 33–58. Routledge.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

Anthropic Team. 2024. Claude 3.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models. Preprint, arXiv:2312.11805.

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

Sheng Wang, Zihao Zhao, Xi Ouyang, Qian Wang, and Dinggang Shen. 2023. Chatcad: Interactive computer-aided diagnosis on medical image using large language models. arXiv preprint arXiv:2302.07257.

Xiyang Wu, Ruiqi Xian, Tianrui Guan, Jing Liang, Souradip Chakraborty, Fuxiao Liu, Brian Sadler, Dinesh Manocha, and Amrit Singh Bedi. 2024. On the safety concerns of deploying llms/vlms in robotics: Highlighting the risks and vulnerabilities. arXiv preprint arXiv:2402.10340.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. arXiv preprint arXiv:2304.12244.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). arXiv preprint arXiv:2309.17421, 9(1):1.

Bohan Zhai, Shijia Yang, Chenfeng Xu, Sheng Shen, Kurt Keutzer, and Manling Li. 2023. Halle-switch: Controlling object hallucination in large vision language models. arXiv e-prints, pages arXiv–2310.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. ArXiv, abs/2309.01219.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592.

A More Experimentation Results

A.1 Synthetic Dataset

Table 4 presents the results of victim LVLMs under three attack strategies using synthetic datasets. GPT-4V-Turbo exhibits the highest robustness to hallucination attacks among all strategies, particularly showing stronger resistance to correctness hallucinations than to consistency hallucinations. Open-source, smaller LVLMs like LLaVA-1.5 and miniGPT4 perform comparably to Gemini and better than Claude. Questions probing the existence of objects are easier to cause hallucination of LVLMs than those probing spatial relations. Question targeting to inserted objects, including the existence questions of abnormal and paired object insertions, contributes to a better hallucination attack success rate than those targeting hypothetical objects, like correlation questions and existence questions for correlated object removal. Hallucination attacks exploiting inconsistencies in responses are more effective for existence questions about inserted objects and spatial relation queries but are less effective for questions about object removal. Results demonstrate that using sequences of questions to probe hallucinations with varying contextual information from the image effectively disrupts the cognitive processing of LVLMs, showing superior results compared to strategies that involve object removal to induce expectation violations in LVLMs.

A.2 Aligned Synthetic Dataset

In an ablation study, we assessed the vulnerability of various LVLMs to three attack strategies using the same synthetic datasets, which incorporate abnormal objects and scene images generated by GPT-4V-Turbo and DALL-E-3. The results, detailed in Table 5, indicate that all LVLMs, except Claude, show a decrease in both MASR and CASR for existence and spatial relation questions, but an increase in attack success rate for correlation questions. This suggests that LVLMs exhibit stronger resistance to hallucinations induced by images from other models than by those they generated themselves, corroborating findings in Section 5.4. GPT-4V-Turbo, in particular, excels in handling paired object insertions. We attribute these differences to the varying priors among LVLMs; a VQA model may perceive an object suggested by another LVLM as less abnormal or correlated, thus reducing the likelihood of hallucinations. Further insights are explored in our ablation study in Section 5.4, where we swap the roles of LVLMs in object prompting and VQA tasks to examine the impact of using different LVLMs for these functions.

A.3 Real-world Dataset

Table 6 displays the results of hallucination attacks using real-world datasets across three strategies. The results indicate that victim LVLMs are more susceptible to hallucination attacks with real-world datasets, showing increased success rates for all metrics compared to those in Table 4. We hypothesize the discrepancy in LVLMs' performance over synthetic and real-world datasets comes from their lack of ability to address the complexity and diversity within the real-world data, which causes its higher vulnerability to our attack strategies when using real-world data.

B Discussion

Across all results discussed in Sections 5.3, 5.4, and Appendix A, we identified key insights into our proposed strategies and LVLMs' resistance to hallucination attacks.

Robust to Absence, Vulnerable to Perturbation. LVLMs are more vulnerable to hallucinations involving object insertions, such as abnormal and paired object insertion strategies, compared to those focused on object absence, like in the correlation object removal strategy. This suggests that attacks leveraging cognitive dissonance through object insertion are more effective than those relying on expectancy violations via object removal.

Robustness to Hallucination Attacks across LVLMs. GPT-4V shows superior resistance to the hallucination attacks we proposed, especially in the MASR metric assessing correctness hallucinations. Gemini slightly outperforms other LVLMs in the CASR metric. Larger models like GPT-4V-Turbo, Gemini Pro Vision, and Claude 3 generally surpass smaller ones such as LLaVA-1.5 and miniGPT4, demonstrating a link between model size and hallucination resistance.

Real-world Data Increases Difficulty. Victim LVLMs show increased vulnerability to hallucination attacks with real-world datasets than synthetic ones. Real-world images generally contain more contextual information than synthetic ones, causing LVLMs to struggle with the added complexity and diversity, thus heightening their vulnerability to hallucination attacks based on real-world data.

Manipulation Strategies	LVLMs	Overall			Existence			Spatial Relation		
		Overall ASR	Overall MASR	Overall CASR	Exi. ASR	Exi. MASR	Exi. CASR	Sp. ASR	Sp. MASR	Sp. CASR
Abnormal Obj. Insertion	GPT-4V-Turbo (Yang et al., 2023)	96.0	80.0	92.5	93.0	62.0	88.5	78.1	71.2	60.6
	Gemini Pro Vision (Team, 2023)	97.0	90.5	90.0	84.5	75.5	68.0	89.1	81.0	73.6
	Claude 3 (Team, 2024)	97.4	90.7	96.0	95.3	81.5	90.7	92.3	79.2	90.8
	LLaVA-1.5 (Liu et al., 2023c)	97.7	94.2	94.0	97.9	87.4	95.6	96.2	83.3	97.6
	miniGPT4 (Zhu et al., 2023)	98.1	95.1	98.0	98.1	89.8	97.7	97.1	89.3	98.2
Paired Obj. Insertion	GPT-4V-Turbo (Yang et al., 2023)	99.5	93.5	97.0	91.5	60.5	86.0	81.7	72.0	58.3
	Gemini Pro Vision (Team, 2023)	100.0	100.0	99.5	99.5	99.5	97.5	85.7	62.3	74.0
	Claude 3 (Team, 2024)	100.0	99.0	99.0	99.0	86.0	98.0	95.5	91.0	91.0
	LLaVA-1.5 (Liu et al., 2023c)	99.7	95.1	98.9	97.6	98.4	94.1	81.8	79.7	72.3
	miniGPT4 (Zhu et al., 2023)	100.0	99.8	100.0	99.1	99.3	99.7	83.9	71.1	75.2
Correlated Obj. Removal	GPT-4V-Turbo (Yang et al., 2023)	93.0	84.0	84.0	69.5	68.5	46.0	85.5	67.6	79.2
	Gemini Pro Vision (Team, 2023)	95.0	92.0	93.0	77.0	77.0	70.5	91.1	83.2	87.4
	Claude 3 (Team, 2024)	99.0	98.0	89.0	92.0	92.0	64.0	88.5	83.3	82.3
	LLaVA-1.5 (Liu et al., 2023c)	97.1	88.9	87.4	70.8	71.4	65.3	87.4	75.3	86.9
	miniGPT4 (Zhu et al., 2023)	96.7	90.1	91.5	72.9	72.7	63.7	86.7	76.4	85.5

Table 4: **Attack Results across all LVLMs with three manipulation strategies on synthetic data.**

Manipulation Strategies	LVLMs	Overall			Existence			Spatial Relation		
		Overall ASR	Overall MASR	Overall CASR	Exi. ASR	Exi. MASR	Exi. CASR	Sp. ASR	Sp. MASR	Sp. CASR
Abnormal Obj. Insertion	GPT-4V-Turbo (Yang et al., 2023)	96.0	80.0	92.5	93.0	62.0	88.5	78.1	71.2	60.6
	Gemini Pro Vision (Team, 2023)	89.5	82.5	76.5	80.5	66.5	64.5	78.8	66.3	60.6
	Claude 3 (Team, 2024)	97.0	93.0	95.0	94.0	82.0	90.0	90.1	84.6	86.8
	LLaVA-1.5 (Liu et al., 2023c)	96.1	79.4	83.3	91.7	70.5	81.4	72.2	68.1	60.4
	miniGPT4 (Zhu et al., 2023)	95.5	72.1	70.9	82.7	61.8	77.2	74.1	70.5	65.8
Paired Obj. Insertion	GPT-4V-Turbo (Yang et al., 2023)	99.5	93.5	97.0	91.5	60.5	86.0	81.7	72.0	58.3
	Gemini Pro Vision (Team, 2023)	100.0	90.5	99.0	83.5	67.0	67.0	78.3	58.3	56.0
	Claude 3 (Team, 2024)	100.0	97.0	100.0	99.0	89.0	99.0	94.2	86.0	90.7
	LLaVA-1.5 (Liu et al., 2023c)	100.0	96.1	98.7	90.3	64.1	87.0	84.4	70.2	57.9
	miniGPT4 (Zhu et al., 2023)	100.0	97.7	99.6	92.7	78.2	89.7	87.8	80.1	67.5
Correlated Obj. Removal	GPT-4V-Turbo (Yang et al., 2023)	93.0	84.0	84.0	69.5	68.5	46.0	85.5	67.6	79.2
	Gemini Pro Vision (Team, 2023)	97.0	94.0	90.5	74.5	74.5	60.5	91.9	83.2	89.0
	Claude 3 (Team, 2024)	100.0	100.0	93.0	94.0	94.0	66.0	90.4	84.3	89.2
	LLaVA-1.5 (Liu et al., 2023c)	98.1	91.2	89.8	70.9	69.9	54.1	87.2	76.1	78.8
	miniGPT4 (Zhu et al., 2023)	97.9	93.5	91.6	78.3	68.1	57.9	89.3	77.4	82.1

Table 5: **Attack Results across all LVLMs with three manipulation strategies using the same synthetic dataset.** This aligned synthetic dataset was created by GPT-4V-Turbo and DALL-E-3, and is used for all victim LVLMs.

Swap Object Prompting and VQA Model Help. According to results in Fig. 4 and Appendix A.2, utilizing different LVLMs to prompt objects for image manipulation and handle VQA tasks reduces hallucinations. This effect is attributed to the varying priors among LVLMs; different models may have different responses to prompted objects for insertion or removal, making some LVLMs more resistant to hallucination cases generated by another model.

C Question Details

Table 7 outlines the details of the questions constructed to probe hallucinations. As outlined in Section 4.3 and 4.4, we employ a series of questions varying in contextual information to explore hallucinations. For questions probing the existence of the target object, we create queries both with and without image-level captions. For those probing the correlation of paired objects, we provide three levels of contextual information: none, the existence of the paired object, and image-level captions. For spatial relation probes, questions utilize the target object's name and descriptive text.

Under each category, we examine conflicts among questions with varying contexts to detect potential consistency in hallucinations.

Manipulation Strategies	LVLMs	Overall			Existence			Spatial Relation		
		Overall ASR	Overall MASR	Overall CASR	Exi. ASR	Exi. MASR	Exi. CASR	Sp. ASR	Sp. MASR	Sp. CASR
Abnormal Obj. Insertion	GPT-4V-Turbo (Yang et al., 2023)	100.0	98.4	98.4	97.6	74.2	92.7	97.5	92.7	89.5
	Gemini Pro Vision (Team, 2023)	100.0	100.0	97.6	97.6	94.3	89.4	94.3	86.2	78.9
	Claude 3 (Team, 2024)	100.0	100.0	100.0	100.0	91.3	100.0	98.4	96.8	98.4
	LLaVA-1.5 (Liu et al., 2023c)	100.0	100.0	98.9	98.6	89.2	92.5	95.9	91.7	92.6
	miniGPT4 (Zhu et al., 2023)	100.0	100.0	97.9	98.0	90.5	92.6	96.1	93.1	87.5
Paired Obj. Insertion	GPT-4V-Turbo (Yang et al., 2023)	100.0	100.0	99.2	99.2	64.5	95.2	100.0	97.6	87.9
	Gemini Pro Vision (Team, 2023)	100.0	99.2	100.0	99.2	84.0	89.6	90.4	84.0	68.8
	Claude 3 (Team, 2024)	100.0	99.2	100.0	97.6	64.3	96.8	99.2	98.4	99.2
	LLaVA-1.5 (Liu et al., 2023c)	99.7	98.5	99.3	94.5	61.8	89.0	97.8	95.6	84.9
	miniGPT4 (Zhu et al., 2023)	100.0	100.0	99.5	99.5	65.7	96.1	99.8	98.6	89.1
Correlated Obj. Removal	GPT-4V-Turbo (Yang et al., 2023)	94.4	88.0	84.0	75.2	72.8	55.2	85.4	73.4	81.5
	Gemini Pro Vision (Team, 2023)	96.8	95.2	92.0	77.6	77.6	68.0	94.2	86.0	89.3
	Claude 3 (Team, 2024)	98.4	98.4	94.4	96.0	95.2	74.6	89.6	88.0	88.8
	LLaVA-1.5 (Liu et al., 2023c)	93.1	97.6	94.6	78.1	73.7	71.1	95.7	89.3	90.1
	miniGPT4 (Zhu et al., 2023)	97.8	96.3	89.1	76.9	73.4	70.4	87.8	74.9	87.5

Table 6: **Attack Results across all LVLMs with three manipulation strategies on a real-world dataset.** The real-world data is created from the Common Objects in Context (COCO) dataset validation set (Lin et al., 2014).

Category	Contextual Info.	Question
Existence	N/A	*Is there a* $\{TargetObjectName\}$ *in this image?*
	Image-level Caption	*We have an image depicting* $\{ImageCaption\}$. *Is there a* $\{TargetObjectName\}$ *in this image?*
Correlation	N/A	*Is there a* $\{ObjectName\}$ *in this image?*
	Paired Obj.	*We have* $\{PairedObjectName\}$ *in this image. Is there a* $\{ObjectName\}$ *in this image?*
	Image-level Caption	*We have an image depicting* $\{ImageCaption\}$. *Is there a* $\{ObjectName\}$ *in this image?*
Spatial Relation	N/A	*Is the* $\{TargetObjectName\}$ $\{spatialrelation\}$ *a/an* $\{ExistingObjectName\}$ *in this image, given their center positions?*
	Obj. Description	*Is the object (*$\{TargetObjectDescription\}$*)* $\{spatialrelation\}$ *a/an* $\{ExistingObjectName\}$ *in this image, given their center positions?*

Table 7: Questions Constructed to Induce Hallucinations

D More Examples

We provide several showcases across all 3 hallucination crafting strategies and all questions covered by AutoHallusion. **Each figure is self-contained for readability**, where we highlight the control pairs, the responses of GPT-4V and LLaVA-1.6, the failures of those models, and the corresponding part of the answers.

Fig. 5 and 6 display cases from the abnormal object insertion strategy. Fig. 5 illustrates both GPT-4V and LLaVA-1.6 inconsistently answering the existence of an inserted object. In Fig.6, only GPT-4V experiences correctness hallucination, while LLaVA-1.6 responds accurately.

Fig. 7 and 8 exhibit cases from the paired object insertion strategy, focusing on the absence of one object paired with an existing object. Fig. 7 shows GPT-4V failing to provide consistent answers across varying contexts, whereas LLaVA-1.6 answers correctly and consistently. In Fig. 8, both models show correctness hallucinations and inconsistency in responses concerning the existence of the paired object.

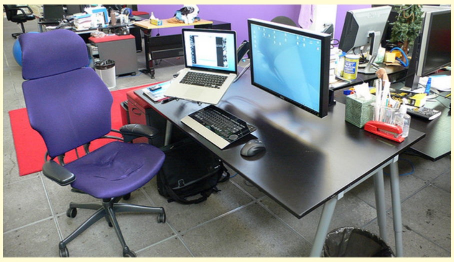
For hallucination cases made by correlated object removal, Fig. 9 shows that both models fail to make correct answers to all questions, while GPT-4V makes wrong answers to both questions and LLaVA-1.6 makes inconsistent answers over questions. The example in Fig. 10 shows that both LVLMs fail to make consistent answers to the spatial relation between the removed object and one of the existing objects in the edited image as they mistakenly assume the existence of the removed object given the contexts presented in the image.

E Failure Case

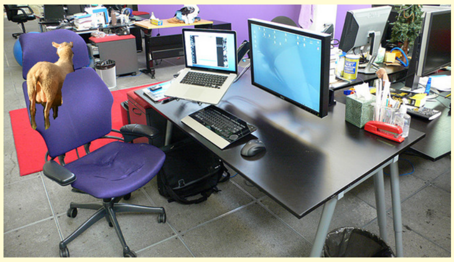
We provide several cases for the failure situation of AUTOHALLUSION we encountered in our experiment. Fig. 11 shows cases when a human could not understand the object being added. Fig. 12 shows cases when LVLMs detect the image manipulation we perform and point out in their answers. Fig. 13 shows cases when the evaluation model fails to provide the correct evaluation answer for the given VQA task.

Abnormal Object Insertion – Existence

Scene: *Messy office workspace*
Detected Objects: *Office swivel chair*, *Open laptop computer*, *Dell computer monitor*, *Office work desk*, *Black laptop backpack*
Inserted Object: *Sheep*
Image Manipulation: *Sheep* is inserted to the image.



Original image



Edited image

Attack: Adding new irrelevant object *Sheep* – Existence
Question: Is there a sheep in this image?
GPT-4V: Yes, there is an image of a sheep that seems to have been digitally added onto a chair in the top half of the image. It looks quite out of place in the office setting depicted in the photo.

Attack: Adding new irrelevant object *Sheep* with image-level caption – Existence
Question: We have an image depicting This image shows a contemporary office workspace featuring several desks cluttered with technological and office equipment. On the left side of the image, there is a vivid scene where a brown goat is digitally superimposed and sitting atop a purple office chair. This humorous addition contrasts with the otherwise ordinary office setting. The workspace is equipped with multiple monitors, laptops, office supplies like pens and staplers, and personal items. The floor is tiled, and the overall appearance suggests a busy, creative, and somewhat messy environment typical of tech or design-oriented workplaces. Is there a sheep in this image?
GPT-4V: No, there isn't a sheep in the image. What you see is a brown goat that is digitally superimposed and sitting on a purple office chair. There are no sheep visible in the picture.

Attack: Adding new irrelevant object *Sheep* – Existence
Question: Is there a sheep in this image?
LLaVA-1.6: Yes, there is a sheep in this image. It appears to be superimposed onto the scene, as it is not a typical setting for a sheep to be found. The sheep is sitting on the back of a purple office chair.

Attack: Adding new irrelevant object *Sheep* with image-level caption – Existence
Question: We have an image depicting This image shows a contemporary office workspace featuring several desks cluttered with technological and office equipment. On the left side of the image, there is a vivid scene where a brown goat is digitally superimposed and sitting atop a purple office chair. This humorous addition contrasts with the otherwise ordinary office setting. The workspace is equipped with multiple monitors, laptops, office supplies like pens and staplers, and personal items. The floor is tiled, and the overall appearance suggests a busy, creative, and somewhat messy environment typical of tech or design-oriented workplaces. Is there a sheep in this image?
LLaVA-1.6: No, there is no sheep in this image. The animal in the image is a goat, not a sheep.

Figure 5: **Hallucination Cases Created by Abnormal Object Insertion:** We highlight hallucination context made by *Correctness*, *Inconsistency*, or potentially mixed.

Abnormal Object Insertion – Spatial Relation

Scene: *Cozy bedroom interior*
Detected Objects: *Blue bed comforter*, *Wooden bookshelf, books*, *Round wall mirror*, *Wooden bedroom dresser*, *Window with view*
Inserted Object: *Stop Sign*
Image Manipulation: *Stop Sign* is inserted to the image.

Original image

Attack: Determine spatial relation between the irrelevant object *Stop Sign* and the existing object: *Round wall mirror*.

Question: Is the stop sign on the right of a/an Round wall mirror. in this image, given their center positions? Where Round wall mirror. is the one that circular mirror with a simple wooden frame, reflecting a room.

GPT-4V: In the image provided, the stop sign is edited into the scene and is located centrally on the windowpane. The round wall mirror, described as a circular mirror with a simple wooden frame, is on the left side of the room above a dresser. Based on their positions, the stop sign is not on the right of the round wall mirror. It is instead to the left of the mirror when viewed from the perspective of someone standing in front of the bed facing the window and the dresser.

Edited image

Attack: Determine spatial relation between the irrelevant object *Stop Sign* and the existing object: *Round wall mirror*.

Question: Is the stop sign on the right of a/an Round wall mirror. in this image, given their center positions? Where Round wall mirror. is the one that circular mirror with a simple wooden frame, reflecting a room.

LLaVA-1.6: Yes, the stop sign is on the right of the Round wall mirror. The mirror is located to the left of the stop sign in the image.

Figure 6: **Hallucination Cases Created by Abnormal Object Insertion:** We highlight hallucination context made by *Correctness*, *Inconsistency*, or potentially mixed.

Paired Object Insertion – Correlated

Scene: *Nighttime urban bus*
Detected Objects: *City transit bus*, *City transit bus*, *Street light pole*, *Urban street tree*, *Bus station canopy*
Inserted Object: *Person*
Hypothetical Object: *Smartphone*
Image Manipulation: *Person* is inserted to the image.

Original image

Attack: Asking about non-existing, correlated object *Smartphone* only – Existence
Question: Is there a/an smartphone in this image?
GPT-4V: No, there is no visible smartphone in the image. The image features a city street scene with a bus and a person running, but there is no clear indication of anyone using or holding a smartphone.

Attack: Asking about non-existing, correlated object *Smartphone* given existing object person – existence
Question: We have person in this image. Is there a/an smartphone in this image as well?
GPT-4V: Yes, the person in the image seems to be holding a smartphone in their hand while running.

Edited image

Attack: Asking about non-existing, correlated object *Smartphone* only – Existence
Question: Is there a/an smartphone in this image?
LLaVA-1.6: No, there is no smartphone visible in this image. The image shows a city street scene with a bus, pedestrians, and a person running.

Attack: Asking about non-existing, correlated object *Smartphone* given existing object person – existence
Question: We have person in this image. Is there a/an smartphone in this image as well?
LLaVA-1.6: No, there is no smartphone visible in this image. The image shows a person running across the street in front of a bus, and there are no electronic devices like smartphones in the scene.

Figure 7: **Hallucination Cases Created by Paired Object Insertion:** We highlight hallucination context made by *Correctness*, *Inconsistency*, or potentially mixed.

Paired Object Insertion – Correlated

Scene: *Foggy lakeside mornings*
Detected Objects: *Wooden fishing boat*, *White swan figurine*, *Foggy lakeside dock*
Inserted Object: *Salt shaker*
Hypothetical Object: *Pepper shaker*
Image Manipulation: *Salt shaker* is inserted to the image.

Original image

Attack: Asking about non-existing, correlated object *Pepper shaker* only – Existence
Question: Is there a/an pepper shaker in this image?
GPT-4V: Yes, there is an image of a pepper shaker overlaying the scene of a tranquil lake, a fishing pier, and a boat. The shaker appears to have been edited into the image and does not naturally belong in the scene.

Attack: Asking about non-existing, correlated object *Pepper shaker* given existing object salt shaker – Existence
Question: We have salt shaker in this image. Is there a/an pepper shaker in this image as well?
GPT-4V: No, there is no pepper shaker in the image. The image only shows a single salt shaker edited into the scene.

Edited image

Attack: Asking about non-existing, correlated object *Pepper shaker* only – Existence
Question: Is there a/an pepper shaker in this image?
LLaVA-1.6: Yes, there is a pepper shaker in the image. It appears to be floating in the air above the water, which is an unusual and surreal element in the scene.

Attack: Asking about non-existing, correlated object *Pepper shaker* given existing object salt shaker – Existence
Question: We have salt shaker in this image. Is there a/an pepper shaker in this image as well?
LLaVA-1.6: No, there is no pepper shaker in this image. The image shows a salt shaker floating in the air above a body of water with a dock and a boat.

Figure 8: **Hallucination Cases Created by Paired Object Insertion:** We highlight hallucination context made by *Correctness*, *Inconsistency*, or potentially mixed.

Correlated Object Removal – Existence

Scene Name: *Modern kitchen interior*
Detected Objects: *Modern kitchen island*, *Glass pendant light*, *Wicker snack basket*, *Round wall clock*
Removed Object: *Bar Stools*
Image Manipulation: *Bar Stools* are replaced with two white towels in the front.

Original image

Attack: Remove an existing object *Bar Stools* – Existence
Question: Is there a/an bar stools in this image?
GPT-4V: Yes, there are bar stools in the image. You can see two of them at the counter with white towels hanging on them.

Attack: Remove an existing object *Bar Stools* with object-level caption – Existence
Question: In this image, We have existing objects including a sleek black kitchen island with an integrated sink and white stools, a clear glass cone-shaped pendant light with a visible filament bulb, A wicker basket filled with various snack items, Gold-rimmed wall clock with large, clear Roman numerals. Is there a/an bar stool in this image?
GPT-4V: Yes, there are white stools visible in the image, positioned at the kitchen island. These could be considered bar stools given their height and setting by the counter.

Edited image

Attack: Remove an existing object *Bar Stools* – Existence
Question: Is there a/an bar stools in this image?
LLaVA-1.6: No, there are no bar stools visible in this image. The image shows a kitchen with a counter, a sink, and some kitchen items on the counter. The lighting is provided by hanging pendant lights, and there's a clock on the wall. The kitchen has a modern design with a mix of black and white elements.

Attack: Remove an existing object *Bar Stools* with object-level caption – Existence
Question: In this image, We have existing objects including a sleek black kitchen island with an integrated sink and white stools, a clear glass cone-shaped pendant light with a visible filament bulb, A wicker basket filled with various snack items, Gold-rimmed wall clock with large, clear Roman numerals. Is there a/an bar stool in this image?
LLaVA-1.6: Yes, there is a bar stool in the image. It is positioned next to the kitchen island, which has an integrated sink. The stool appears to be white, matching the overall color scheme of the kitchen.

Figure 9: **Hallucination Cases Created by Correlated Object Removal:** We highlight hallucination context made by *Correctness*, *Inconsistency*, or potentially mixed.

Correlated Object Removal – Spatial Relation

Scene Name: *Pizza, wine, street-view*
Detected Objects: *Seafood pizza slice*, *Banana pudding dessert*, *Pizza cutting knife*, *Chef's hat*
Removed Object: *Wine Glasses*
Image Manipulation: *Wine Glasses* are removed from the image.

Original image

Edited image

Attack: Remove an existing object *Wine Glasses* – Existence
Question: Is there a/an wine glasses in this image?
GPT-4V: There are no wine glasses visible in the image. The scene primarily shows two pizzas on plates, various condiments in bowls, and the interior of a restaurant. There is also a slice of pizza being lifted by a person at the table.

Attack: Determine the spatial relation between the removed object *Wine Glasses* and the existing object: *Seafood pizza slice*. The model should insist no.
Question: Is the wine glasses on the left of a/an Seafood pizza slice. in this image, given their center positions? The Seafood pizza slice is the one that a slice of seafood pizza topped with shrimp and vegetables.
GPT-4V: From the image provided, we can see two pizza pans: one in the center and one towards the right. The slice of pizza on the right pan appears to be topped with shrimp and vegetables, characteristic of a seafood pizza. As for the wine glasses, they are indeed positioned to the left of the slice of seafood pizza, relative to the image's viewpoint. They sit near the left edge of the table.

Attack: Remove an existing object *Wine Glasses* – Existence
Question: Is there a/an wine glasses in this image?
LLaVA-1.6: No, there are no wine glasses in this image. The table is set with plates of food, including pizza, and various bowls, but there are no wine glasses visible.

Attack: Determine the spatial relation between the removed object *Wine Glasses* and the existing object: *Seafood pizza slice*. The model should insist no.
Question: Is the wine glasses on the left of a/an Seafood pizza slice. in this image, given their center positions? The Seafood pizza slice is the one that a slice of seafood pizza topped with shrimp and vegetables.
LLaVA-1.6: Yes, the wine glasses are on the left of the Seafood pizza slice in this image. The Seafood pizza slice is the one that has a slice of seafood pizza topped with shrimp and vegetables.

Figure 10: **Hallucination Cases Created by Correlated Object Removal:** We highlight hallucination context made by *Correctness*, *Inconsistency*, or potentially mixed.

Failure Case: Non-Perceivable Objects Prompted

Scene Name: *Hiking, horseback, wilderness*
Detected Objects: *Man riding horse*, *Large gray rock*, *Riding horse forward*, *Red bandana, glasses*
Inserted Object: *Refrigerator*
Image Manipulation: *Refrigerator* is inserted into the image.

Original image

Edited image

Attack: Adding new irrelevant object *Refrigerator* – Existence
Question: Is there a/an refrigerator in this image?
GPT-4V: No, there is no refrigerator in this image. The image shows a man wearing a bandana looking at another man riding a horse, which is carrying what appear to be large wooden boxes or panels. They are on a rocky trail in a forested area.
LLaVA-1.6: No, there is no refrigerator in this image. The image shows a person riding a horse, with another person walking alongside, and a third person looking back at the camera. They are in a natural setting with trees and rocks.

Scene Name: *Fishing boats, marshland*
Detected Objects: *Harbor sailing boats*, *Marshland with egrets*, *Blurred swan figure*, *Fishing boat masts*
Inserted Object: *Airplane*
Image Manipulation: *Airplane* is inserted into the image.

Original image

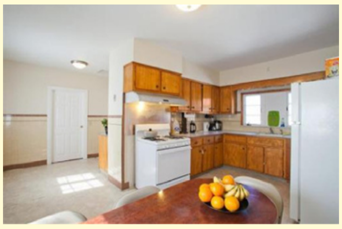
Edited image

Attack: Adding new irrelevant object *Airplane* – Existence
Question: Is there a/an airplane in this image?
GPT-4V: Yes, there is an airplane in the image. It is flying in the upper left part of the photo, above the harbor area and the grassy field.
LLaVA-1.6: Yes, there is an airplane in the image. It appears to be flying in the sky above the landscape.

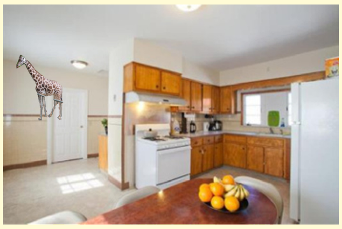
Figure 11: **Failure Case: Non-Perceivable Objects Prompted.**

Failure Case: When LVLMs Detect the Edition

Scene Name: *Bright, tidy kitchen*
Detected Objects: *White kitchen dishwasher*, *White vertical board*, *Wooden kitchen cabinets*, *White kitchen dishwasher*, *Bowl of oranges*
Inserted Object: *Giraffe*
Image Manipulation: *Giraffe* is inserted into the image.



Original image



Edited image

Attack: Adding new irrelevant object *Giraffe* – Existence
Question: Is there a/an giraffe in this image?
GPT-4V: There isn't an actual giraffe in the image. The image includes a graphical or edited-in representation of a giraffe in a kitchen setting, which is a fictional and humorous element because giraffes are not indoor animals nor they are typically found in such environments.
LLaVA-1.6: Yes, there is a giraffe in this image. It appears to be a digitally added or edited element, as it is not possible for a giraffe to be in a kitchen setting like this.

Scene Name: *Mountain cottage*
Detected Objects: *Burning wooden logs*, *Snowshoes hanging outside*, *Snowy pine tree*, *Enclosed wood fireplace*
Inserted Object: *Palm tree*
Image Manipulation: *Palm tree* is inserted into the image.

Original image

Edited image

Attack: Adding new irrelevant object *Palm tree* – Existence
Question: Is there a/an Palm tree in this image?
GPT-4V: Yes, there is an image of a palm tree in this picture. It appears to have been artificially placed or edited into the snowy winter scene, as palm trees are not naturally found in such cold, snowy environments. The palm tree is located near the center of the image and is highlighted with a blue neon-like outline.
LLaVA-1.6: Yes, there is a palm tree in this image. It's quite unusual to see a palm tree in a snowy, mountainous setting, which makes the image quite surreal.

Figure 12: **Failure Case: When LVLMs Detect the Edition.**

Failure Case: When Evaluation Model Fails

Scene Name: *Children petting goat*
Detected Objects: *Black and white goat*, *Floral summer dress*, *Black domestic goat*, *Illuminated digital keypad*
Inserted Object: *Microwave*
Image Manipulation: *Microwave* is inserted into the image.

Original image

Edited image

Attack: Adding new irrelevant object *Microwave* – Existence
Question: Is there a/an microwave in this image?
GPT-4V: No, there is not a microwave in the original context of the image. The photo shows two young girls interacting with goats in an outdoor setting, presumably at an animal pen or a small farm. The microwave appears to be an edited addition to the image.
Ground Truth: There is a microwave.
Evaluation Result: Not same.

Scene Name: *Cafe dessert table*
Detected Objects: *Layered cream cake*, *Stainless steel forks*
Removed Object: *Glass*
Image Manipulation: One *Glass* is removed from the image, while there's another *Glass* is the edited image escaped from the object detection and removal model.

Original image

Edited image

Attack: Remove an existing object *Glass* – Existence
Question: Is there a/an glass in this image?
GPT-4V: Yes, there is a glass in the image. It is located at the lower left portion of the picture, next to the piece of cake on the table.
Ground Truth: There is no glass.
Evaluation Result: Not same.

Figure 13: **Failure Case: When Evaluation Model Fails.**