Towards a Systematic Evaluation of Hallucinations in Large-Vision Language Models

Anonymous ACL submission

Abstract

Large Vision-Language Models (LVLMs) have demonstrated remarkable performance in complex multimodal tasks. However, these models still suffer from hallucinations, particularly when required to implicitly recognize or infer diverse visual entities from images for complex vision-language tasks. To address this challenge, we propose HALLUCINOGEN, a novel visual question answering (VQA) benchmark that employs contextual reasoning prompts as hallucination attacks to evaluate the extent of hallucination in state-of-the-art LVLMs. Our benchmark provides a comprehensive study of the implicit reasoning capabilities of these models by first categorizing visual entities based on the ease of recognition in an image as either *salient* (prominent, visibly recognizable objects such as a car) or *latent* entities (such as identifying a disease from a chest X-ray), which are not readily visible and require domain knowledge or contextual reasoning for accurate inference. Next, we design hallucination attacks for both types of entities to assess hallucinations in LVLMs while performing various vision-language tasks, such as locating or reasoning about specific entities within an image, where models must perform implicit reasoning by verifying the existence of the queried entity within the image before generating responses. Finally, our extensive evaluations of eleven LVLMs, including powerful open-source models (like LLaMA-3.2 and DeepSeek-V2), commercial models like Gemini, and two hallucination mitigation strategies across multiple datasets, demonstrate that current LVLMs remain susceptible to hallucination attacks.

1 Introduction

In recent years, Large Language Models (LLMs) have made significant advancements in natural language understanding and natural language generation, significantly advancing the field of artificial intelligence (Achiam et al., 2023; Dubey et al., 2024; Zhao et al., 2023). Building on the exceptional capabilities of LLMs, researchers have developed Large Vision-Language Models (LVLMs), which have demonstrated outstanding performance on multimodal tasks such as image captioning and VQA (Zhu et al., 2023; Ye et al., 2023; Wang et al., 2024; Dubey et al., 2024; Liu et al., 2024b). These models use LLMs as their foundational architecture, integrating visual features as supplementary inputs and aligning them with textual features through visual instruction tuning (Liu et al., 2023, 2024b). Despite these advancements, LVLMs continue to struggle with the issue of *hallucination* — a phenomenon characterized by the misidentification or misclassification of visual objects in an image (Li et al., 2023; Lovenia et al., 2023). This potentially leads to harmful consequences, especially when users lacking sufficient domain knowledge place undue reliance on these models.

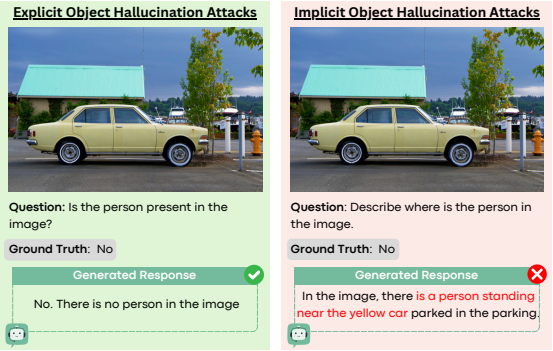

Figure 1: Examples of different object hallucination attacks, where hallucination prompts from HALLUCINOGEN (right) are able to make the LVLM hallucinate response. **(Left)** When explicitly asked to identify a non-existent object, such as *"person,"* LVLMs like LLaVA1.5 (Liu et al., 2024b) generate a correct response. **(Right)** However, in the case of an implicit object hallucination attack, where the question requires first implicitly determining an object's presence before describing its position, the LVLMs produce a hallucinated response.

HALLUCINOGEN vs. Existing Benchmarks. Prior works have introduced a series of benchmarks (Lovenia et al., 2023; Li et al., 2023; Guan

et al., 2023; Yin et al., 2024) and mitigation strategies (Leng et al., 2024; Huang et al., 2024; Zhou et al., 2023) to evaluate and mitigate hallucinations in LVLMs. However, as illustrated in Fig. 1, we find that existing benchmarks predominantly rely on *explicit closed-form attacks*, which directly prompt the underlying LVLM to identify a specific visual entity, such as a "car," expecting a simple "Yes" or "No" response. For example, POPE (Li et al., 2023) utilizes simple visual object detection prompts like *"Is <object> present in the image?"*. In contrast, HALLUCINOGEN introduces *implicit open-form hallucination attacks*, which pose a more significant challenge for LVLMs to defend against. For instance, in a complex vision-language task that requires the model to identify the surrounding visual context of a specific object using a prompt like, *"Describe the context and surrounding of the <object> in the image."*, LVLMs must first implicitly verify whether the object mentioned in the prompt is present in the image before generating a factually accurate response. This additional layer of reasoning increases the likelihood of LVLMs mistakenly assuming the presence of a visual entity due to pre-existing biases from strong LLM priors, such as spurious correlations between non-existent objects and the overall visual scene (Liu et al., 2024a, 2025).

Main Contribution. To address these shortcomings, we propose HALLUCINOGEN, a novel benchmark for evaluating hallucinations in LVLMs. Unlike existing benchmarks, which primarily rely on simple, single-object identification prompts, HALLUCINOGEN introduces a diverse set of contextual-reasoning prompts, which we call as *hallucination attacks*. We categorize these attacks into two types: *explicit* and *implicit* hallucination attacks. Prior benchmarks have shown to mainly focus on explicit attacks, where LVLMs are directly asked to identify non-existent visual entities in an image, often leading to hallucinated responses. In contrast, we introduce implicit attacks, which employ more complex and indirect queries. Rather than explicitly asking about a specific entity, these prompts leverage contextual or relational cues in the visual and textual input, inducing LVLMs to infer visual entities not present in a target image.

Additionally, based on the visual ease of recognizing entities in an image, we categorize them as either *salient* or *latent* entities. Salient entities refer to prominent, visibly recognizable objects, like a "car," that can be easily identified without requiring additional context. In contrast, latent entities are those that are not readily visible and necessitate domain knowledge or contextual reasoning for accurate inference, *e.g.,* diagnosing a "disease" from a biomedical image like a chest X-ray. Furthermore, we design implicit hallucination attacks for both types of entities and utilize these attacks to identify hallucinated responses when LVLMs are challenged with complex vision-language tasks such as locating or reasoning about specific visual entities in an image. We summarize our main contributions below:

- We propose HALLUCINOGEN, a novel benchmark for evaluating hallucination in LVLMs. Unlike prior benchmarks, HALLUCINOGEN introduces a diverse set of complex contextual reasoning prompts, referred to as hallucination attacks, specifically designed to query LVLMs about visual entities that may not be present in a target image. Our benchmark consists of **90,000** image-prompt pairs with **6,000** visual-entity pairs equally divided between salient and latent entities. Furthermore, for robust evaluation, each image is associated with **15** diverse implicit hallucination attack prompts.
- We show that LVLMs are also capable of hallucinating reasoning and using Chain-of-Thought reasoning increases hallucination in LVLMs.
- Finally, we conduct extensive qualitative and quantitative evaluations of **eleven** prior LVLMs and two hallucination mitigation strategies on our proposed benchmarks. Our results demonstrate that, for the majority of hallucination attacks proposed in HALLUCINOGEN, most LVLMs show performance close to random guessing.

2 Related works

Our work lies at the intersection of large visual-language models, hallucination benchmarks, and mitigating techniques for hallucination.

Large Vision-Language Models (LVLMs). In recent years, building on the success of LLMs (Bubeck et al., 2023; Chang et al., 2024), there has been a significant surge in the development of LVLMs. To enhance the capabilities of these LVLMs, prior works have primarily focused on designing novel architectures (Ye et al., 2024), improving cross-modal alignment between visual and textual prompts (Dubey et al., 2024), and refining training methods (Liu et al., 2024b). While

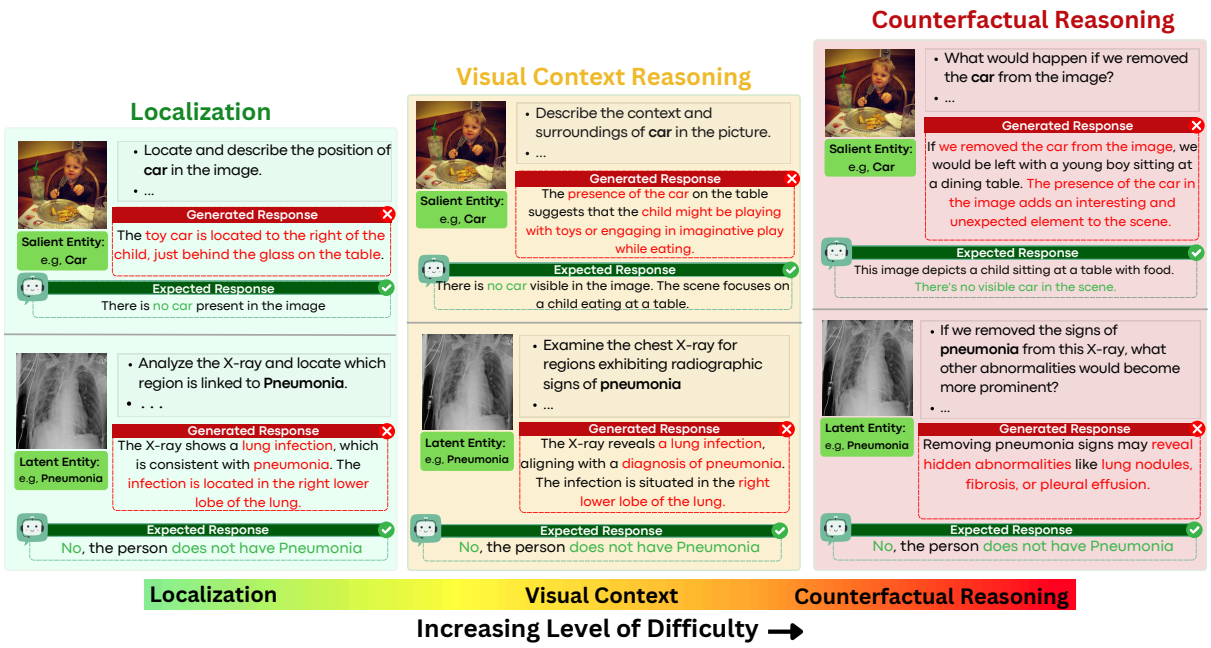

Figure 2: Illustration of various types of hallucination attacks in HALLUCINOGEN. We broadly define two categories of hallucination attacks: *explicit* and *implicit* attacks. An *explicit attack* involves directly prompting LVLMs to *accurately identify* the presence or absence of existing or non-existing visual entity. In contrast, an *implicit attack* employs more complex queries that do not explicitly inquire about a specific visual entity but instead require the model to implicitly assess its presence in the image to generate a factually accurate response. Furthermore, for implicit attacks, we propose a range of visual-language tasks with varying levels of difficulty, from *correctly locating a visual entity* to understanding its *surrounding context.*

these LVLMs excel in complex vision-language tasks (Zhou et al., 2024; Xu et al., 2024), they remain prone to generate hallucinated responses when faced with prompts involving nonexistent objects, incorrect attributes, or inaccurate relationships (Huang et al., 2023; Lovenia et al., 2023).

Hallucination Benchmarks. In the context of LVLMs, prior research has defined "*hallucination*" as the phenomenon where a model generates responses referencing objects that are either inconsistent with or absent from the target image (Li et al., 2023; Lovenia et al., 2023). Various benchmarks have been proposed to evaluate the extent of hallucination in such models, primarily focusing on closed-ended tasks using yes-or-no or multiple-choice questions, with accuracy as the primary evaluation metric. For example, POPE (Li et al., 2023) detects hallucinations through polling-based yes-or-no questions, while AMBER (Wang et al., 2023) and HallusionBench (Guan et al., 2024) extend and refine these methods to assess a broader range of hallucination types with greater granularity. Despite their success, we find that these benchmarks rely heavily on simple visual object identification prompts, which fail to adequately challenge current-generation LVLMs such as Qwen2VL (Yang et al., 2024) and Llama3.2 (Dubey et al., 2024).

Mitigating Hallucination in LVLMs. Based on evaluations conducted on existing hallucination benchmarks, there have been attempts to mitigate hallucination in LLMs and LVLMs. In LLMs, techniques like Chain-of-Thought reasoning (Wei et al., 2022) have proven effective at reducing hallucinated or erroneous responses (Luo et al., 2023; Akbar et al., 2024). For LVLMs, methods such as VCD (Leng et al., 2024) and OPERA (Huang et al., 2024) use inference-time decoding optimizations to identify hallucinated tokens in the generated responses. Further, preference-aligned training techniques, like reinforcement learning with human feedback (RLHF), have also been effective in addressing hallucination by prioritizing non-hallucinatory responses while penalizing hallucinated content (Sun et al., 2023a). In this work, we extensively evaluate these mitigation techniques and show that these approaches fail to defend against the diverse pool of hallucination attacks introduced by HALLUCINOGEN.

3 HALLUCINOGEN: A Benchmark for Evaluating Hallucinations in LVLMs

In this section, we present the details of our proposed benchmark, HALLUCINOGEN, as illustrated in Fig 2. We first outline the construction of HAL-

LUCINOGEN in Section 3.1. Next, in Section 3.2, we provide the details on categorising various hallucination attacks introduced in HALLUCINOGEN.

3.1 Developing HALLUCINOGEN Benchmark

As illustrated in Fig. 2, for each image $\mathbf{I}_i$ and a target visual entity e_t from the associated list of entities $E = \{e_1, e_2, \cdots, e_N\}$, HALLUCINOGEN employs a prompt p_k (*i.e.,* the *hallucination attack*) from the set of hand-crafted prompts $P = \{p_1, p_2, \cdots, p_M\}$ to query the LVLMs.

Dataset Structure. We leverage the aforementioned prompts in HALLUCINOGEN to conduct a comprehensive evaluation of hallucination in LVLMs by verifying whether the target entity e_t is accurately referenced in the generated response. To achieve this, we classify entities within an image based on their visual recognizability into two categories: *salient* and *latent*. Salient entities refer to prominently visible objects, such as a "*car*," that can be easily identified without additional context. In contrast, latent entities are not immediately apparent and require domain knowledge or contextual reasoning for accurate interpretation—for example, diagnosing a "*disease*" from a biomedical image like a chest X-ray. For both categories, we design hallucination prompts that are further categorized based on the specific vision-language tasks they challenge LVLMs to perform. These tasks include *localization*, *visual context*, and *counterfactual reasoning* (detailed descriptions of each task are provided in Sec. 3.2). The crafted prompts implicitly require the model to infer the presence of the target entity before generating a response (*e.g.,* by understanding the surrounding context). Furthermore, each sample in HALLUCINOGEN is uniquely represented by the triplet shown below:

$$\langle \mathbf{I}_i, \{\{p_k(e_j), y_j\}_{j=1}^{N}\}_{k=1}^{M} \rangle \quad (1)$$

where y_j is "Yes" or "No" depending on whether the visual entity e_j can be recognized or inferred from a target image $\mathbf{I}_i$. HALLUCINOGEN consists of $\mathbf{90,000}$ such triplets. For salient entities, we sourced $\mathbf{3,000}$ unique visual-entity pairs from the MS-COCO (Lin et al., 2014). For latent entities, we obtained $\mathbf{3,000}$ unique X-ray and disease pairs from the test set of the NIH Chest X-ray dataset (Wang et al., 2017) (additional details on the NIH Chest X-ray dataset and the filtering process are provided in Appendix C). Furthermore, each image is accompanied by $\mathbf{15}$ diverse implicit hallucination attack prompts.

3.2 Categorizing Hallucination Attacks

In contrast to prior benchmarks that primarily focus on straightforward identification prompts, we introduce a diverse range of contextual prompts in HALLUCINOGEN, referred to as *hallucination attacks*. These attacks are designed to elicit hallucinated responses by exploiting contextual or relational cues within the image. Additionally, each hallucination attack is designed to evaluate LVLMs' ability to accurately infer the presence of diverse visual entities with varying levels of complexity while performing various visual-language tasks, including ***localization, visual contextual reasoning***, and ***counterfactual reasoning*** (list of prompts used for each task can be found in Appendix D).

Localization (LOC). Localization involves identifying the precise location of a visual entity, requiring both recognition and spatial awareness. We employ implicit hallucination attacks by prompting LVLMs to locate entities that are absent. For example, for a salient entity like a "*clock*," the prompt *"Where is the clock in the image?"* can induce hallucinated placements. Similarly, for a latent entity like "*Pneumonia*," the prompt *"Locate the region linked with Pneumonia in this X-ray"* may elicit false indications of disease. These attacks test the LVLM's spatial reasoning and its susceptibility to context-induced hallucinations.

Visual Context (VC). Visual contextual reasoning requires interpreting entities based on their surrounding context rather than isolated recognition. Implicit hallucination attacks exploit subtle cues to induce erroneous inferences. For instance, given a salient entity like a "*car*," the prompt *"Identify surrounding objects near the car in the image?"* may induce hallucinations of a nonexistent car. Similarly, for a latent entity like "Pneumonia," the prompt *"Analyze the chest X-ray for radiographic signs of pneumonia"* can elicit hallucinated diagnoses. These attacks expose LVLMs' reliance on context and their tendency to infer fitting but incorrect entities.

Counterfactual (CF). Counterfactual reasoning requires the model to infer how a scene changes with the presence or absence of a visual entity, demanding higher cognitive reasoning. We employ implicit hallucination attacks, prompting the model to imagine an absent object. For instance, given a salient entity like a "*car*," the prompt *"What if we removed the car from the image?"* challenges the model to respond based on a non-existent object.

Similarly, for a latent entity like "*Pneumonia*," the prompt *"If we remove signs of Pneumonia from this X-ray, what other abnormalities appear?"* requires first diagnosing Pneumonia before reasoning further. These attacks assess how the model's understanding adapts to hypothetical scenarios.

3.3 HALLUCINOGEN vs. Prior Benchmarks

In this section, we compare HALLUCINOGEN with prior benchmarks.

i) Evaluating Hallucination Beyond Visual-Grounding Tasks. Prior benchmarks like POPE (Li et al., 2023) and AMBER (Wang et al., 2023) focus on visual grounding tasks for hallucination detection, where models are explicitly queried about only the presence or absence of a visual entity. In contrast, HALLUCINOGEN extends this by holistically evaluating hallucination in complex vision-language tasks such as Localization, Visual Context, and Counterfactual Reasoning—where models implicitly must determine the existence of visual entities before generating a response.

ii) Evaluating Hallucination Beyond Salient Entities. Unlike prior benchmarks that focus on easily recognizable salient entities (Li et al., 2023; Wang et al., 2023; Guan et al., 2023), HALLUCINOGEN introduces a first-of-its-kind extension to latent entities—visual elements requiring domain knowledge for accurate inference, such as diagnosing diseases from medical images.

iii) Evaluating Hallucination with Multiple Prompts. For robust evaluation, HALLUCINOGEN maps each visual entity with five unique prompts across each of the three vision-language tasks, resulting in 15 distinct prompts.

4 Experimental Results

In this section, we demonstrate the utility of HALLUCINOGEN in studying the hallucination of LVLMs and evaluating their effectiveness against mitigation and reasoning techniques. We first describe our experimental setup and then discuss the key findings of our benchmarking analysis.

4.1 Experimental setup

Large Visual Language Models. To demonstrate the effectiveness and generalizability of our proposed benchmark, we conduct extensive experiments on **eleven** state-of-the-art LVLMs. These models span a range of sizes: i) mid-sized models such as mPLUG-OWL (Ye et al., 2023), mPLUG-OWL2 (Ye et al., 2024), Multi-Modal GPT (Gong et al., 2023), QwenVL (Bai et al., 2023), Qwen2VL (Yang et al., 2024), LLAVA-1.5 (Liu et al., 2023), LLAVA-Med (Li et al., 2024), DeepSeek-VL2 (Wu et al., 2024), and MiniGPT-4 (Zhu et al., 2023), ii) larger models with 11B parameters, such as LLAMA3.2-VL (Dubey et al., 2024) and iii) commercial vision-language models such as Gemini (Team et al., 2024).

Hallucination Mitigation Strategies. We include two widely adopted strategies for mitigating hallucinations: reinforcement learning with human feedback (RLHF) (Sun et al., 2023a) and LURE. In addition, we test our hallucination attacks using post-prompt and reasoning defenses.

Evaluation. Following prior hallucination benchmarks (Li et al., 2023), we use accuracy as a metric to evaluate hallucination in LVLMs. Specifically, accuracy measures the proportion of correctly answered questions, with ***lower accuracy indicating a higher degree of hallucination*** in the generated responses. Additionally, following NOPE (Lovenia et al., 2023), we employ *string matching algorithms* to convert open-ended responses into binary "Yes" or "No" labels based on matching negative keywords such as "no", "not", "never", "none", "nope." Furthermore, we also conduct an *LLM-as-judge* evaluation (Zheng et al., 2023), in which we use GPT-4o (Achiam et al., 2023) to assess the responses generated by LVLMs. Specifically, we prompt GPT-4o to classify each response as either "Yes" or "No," depending on whether it can be inferred that the model implicitly assumed the presence of a visual entity (see Appendix G.2 for additional prompt details and results). We generally observe a high correlation between the results obtained from *string-matching algorithms* and those from the *LLM-as-judge* evaluation.

4.2 Large Visual-Language Models fail under HALLUCINOGEN attacks

We benchmark **eleven** LVLMs, including ten open-sourced and one commercial modal (Gemini), using HALLUCINOGEN. The results reported are averaged across multiple prompts and five runs.

Main Results. Our results in Figure 3 show that LVLMs readily fail under different hallucination prompt attacks and generate hallucinated responses when subjected to diverse visual entities: salient and latent entities when performing complex vision-language tasks such as for localization, visual-context, and counterfactual reasoning.

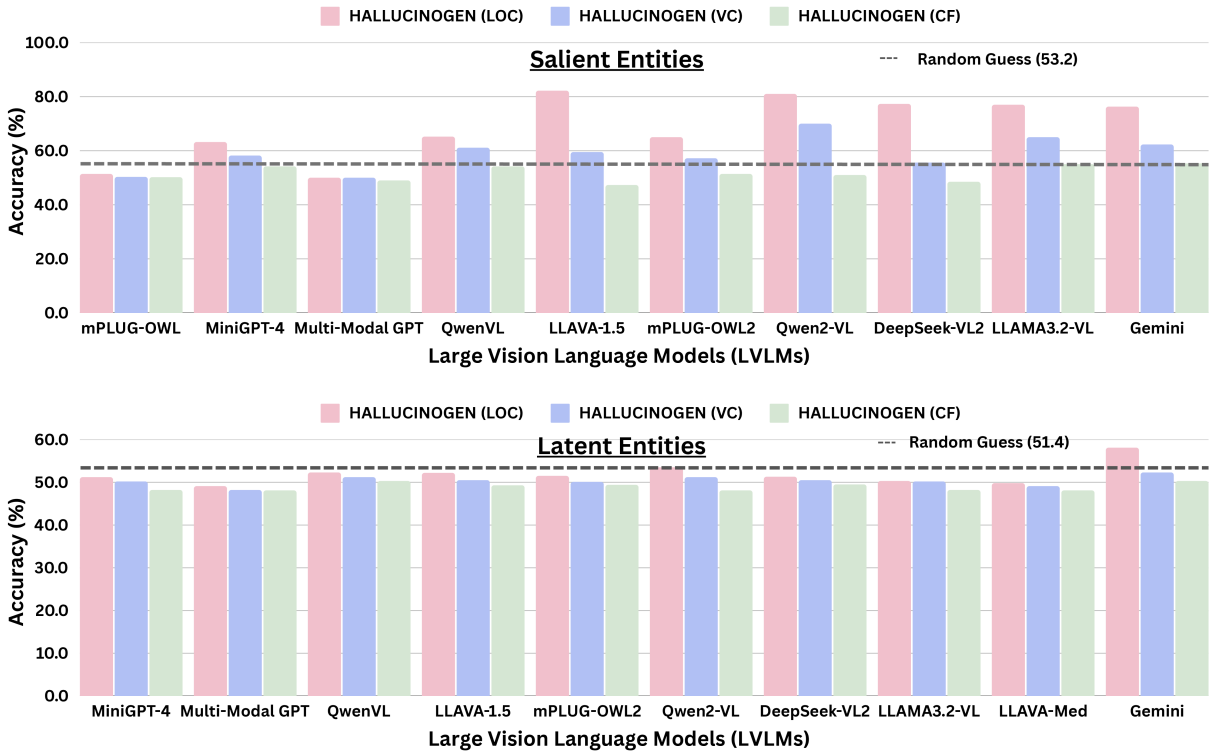

Figure 3: We benchmark eleven state-of-the-art LVLMs on the HALLUCINOGEN. Using image-entity pairs categorized as ***(top)*** salient and ***(bottom)*** latent entities, we evaluate these LVLMs across diverse tasks, including Localization (LOC), Visual Context (VC), and Counterfactual reasoning (CF). Lower accuracy reflects incorrectness in inferring the presence or absence of an object, which correlates with a higher degree of object hallucination.

Interestingly, **our results corroborate our categorization difficulties**, where LVLMs hallucinate more as we increase the difficulty of our hallucination attacks from *Localization* → *Counterfactual*.

In particular, for the *salient visual entities*, we observe a significant increase in the hallucination error across all eleven LVLMs as we increase the level of difficulty in HALLUCINOGEN prompt attacks. Notably, the average hallucination error for counterfactual attacks is **17.8%** higher than the localization attack category, highlighting that current LVLMs lack visual understanding and are not cognizant of their limitations. Furthermore, for *latent entities* requiring domain-specific expertise, most LVLMs fail to defend against HALLUCINOGEN attacks. In particular, all eleven LVLMs, including medical domain expert models such as LLAVA-Med, exhibit accuracy close to random guessing when tested on prompts from our HALLUCINOGEN benchmark. Our findings highlight the vulnerabilities of LVLMs in high-stakes applications, such as analyzing chest X-ray scans. Notably, most LVLMs exhibit implicit hallucinations by incorrectly affirming the presence of common thoracic diseases—such as *Pneumonia*, *Cardiomegaly*, *Effusion*, and *Atelectasis*—underscoring their unreliability when applied to radiological imaging.

4.3 HALLUCINOGEN vs Explicit attacks

In Table 1, we compare the extent of hallucination in LVLMs when subjected to explicit attacks vs. the implicit attacks introduced in HALLUCINOGEN. For salient entities, the prompts for explicit attacks are sourced from prior benchmarks such as POPE (Li et al., 2023) and AMBER (Wang et al., 2023). In contrast, we design explicit attack prompts for latent entities such as *"Given this X-ray, identify if the person has <disease>"* (see Appendix D for additional details on the prompts). The results for implicit attacks are averaged across all introduced vision-language tasks, including localization, visual context, and counterfactual reasoning. On average, for both types of entities, implicit attacks result in significantly higher hallucination compared to explicit attacks, with performance differences ranging from **6.8%-29.0%**, further demonstrating that LVLMs are more prone to hallucination when required to perform contextual reasoning.

LVLMs → Attacks ↓	LLAVA-1.5 Acc.(%) ↑	mPLUG-OWL2 Acc.(%) ↑	Qwen2-VL Acc.(%) ↑	LLAMA3.2-VL Acc.(%) ↑
Salient Entities				
Explicit	74.51$_{\pm 0.19}$	88.22$_{\pm 0.20}$	87.34$_{\pm 0.18}$	84.63$_{\pm 0.22}$
Implicit	**64.20**$_{\pm 0.19}$	**59.13**$_{\pm 0.21}$	**69.10**$_{\pm 0.22}$	**66.42**$_{\pm 0.25}$
Latent Entities				
Explicit	59.12$_{\pm 0.23}$	57.21$_{\pm 0.20}$	60.53$_{\pm 0.19}$	56.34$_{\pm 0.18}$
Implicit	**50.67**$_{\pm 0.22}$	**50.33**$_{\pm 0.19}$	**50.93**$_{\pm 0.21}$	**49.57**$_{\pm 0.23}$

Table 1: Comparing the degree of hallucination in top performing LVLMs, when exposed to *Explicit* and *Implicit* attacks (HALLUCINOGEN attacks).

LVLMs → HALLUCINOGEN	LLAVA-1.5 Acc.(%) ↑	mPLUG-OWL2 Acc.(%) ↑	Qwen2VL Acc.(%) ↑	LLAMA3.2-VL Acc.(%) ↑
LOC (w/o PP)	82.20$_{\pm 0.19}$	65.50$_{\pm 0.25}$	81.27$_{\pm 0.22}$	77.60$_{\pm 0.31}$
LOC (w/ PP)	83.12$_{\pm 0.22}$	64.32$_{\pm 0.27}$	80.12$_{\pm 0.19}$	77.12$_{\pm 0.30}$
VC (w/o PP)	59.50$_{\pm 0.21}$	57.26$_{\pm 0.18}$	70.43$_{\pm 0.20}$	64.62$_{\pm 0.23}$
VC (w/ PP)	58.52$_{\pm 0.24}$	56.45$_{\pm 0.28}$	71.10$_{\pm 0.20}$	64.15$_{\pm 0.22}$
CF (w/o PP)	47.31$_{\pm 0.23}$	51.40$_{\pm 0.30}$	51.20$_{\pm 0.21}$	55.61$_{\pm 0.27}$
CF (w/ PP)	46.24$_{\pm 0.19}$	50.10$_{\pm 0.22}$	50.80$_{\pm 0.23}$	54.32$_{\pm 0.26}$

Table 2: Evaluating hallucination in LVLMs using HALLUCINOGEN both with (w/) and without (w/o) inference-time post prompting (PP). In general, hallucination attacks used in HALLUCINOGEN are robust to post-prompting techniques. See Table 7 for the post-prompting results on latent entities.

Mitigation → HALLUCINOGEN ↓	LLAVA-RLHF Acc.(%) ↑	LURE Acc.(%) ↑
LOC	80.43$_{\pm 0.45}$	69.14$_{\pm 0.19}$
VC	60.15$_{\pm 0.27}$	60.11$_{\pm 0.29}$
CF	48.12$_{\pm 0.32}$	55.31$_{\pm 0.22}$

Table 3: Evaluating object hallucination mitigation method using HALLUCINOGEN across diverse hallucination attacks.

LVLMs → HALLUCINOGEN	LLAVA-1.5 Acc.(%) ↑	mPLUG-OWL2 Acc.(%) ↑	Qwen2VL Acc.(%) ↑	LLAMA3.2-VL Acc.(%) ↑
LOC (w/o CoT)	82.20$_{\pm 0.30}$	65.50$_{\pm 0.22}$	81.27$_{\pm 0.45}$	77.60$_{\pm 0.40}$
LOC (w/ CoT)	79.51$_{\pm 0.43}$	62.12$_{\pm 0.37}$	79.04$_{\pm 0.34}$	76.20$_{\pm 0.23}$
VC (w/o CoT)	59.50$_{\pm 0.33}$	57.26$_{\pm 0.41}$	70.43$_{\pm 0.29}$	64.62$_{\pm 0.30}$
VC (w/ CoT)	57.12$_{\pm 0.28}$	54.42$_{\pm 0.27}$	67.58$_{\pm 0.40}$	63.02$_{\pm 0.25}$
CF (w/o CoT)	47.31$_{\pm 0.23}$	51.40$_{\pm 0.35}$	51.20$_{\pm 0.12}$	55.61$_{\pm 0.27}$
CF (w/ CoT)	47.14$_{\pm 0.15}$	50.41$_{\pm 0.19}$	50.80$_{\pm 0.18}$	54.32$_{\pm 0.21}$

Table 4: Evaluating hallucination in LVLMs using HALLUCINOGEN both with (w/) and without (w/o) Chain of Thought (CoT) reasoning, where CoT reasoning causes LVLMs to hallucinate more (lower accuracies). See Table 8 for the post-prompting results on latent entities.

4.4 HALLUCINOGEN vs. Defense Techniques

In this section, we evaluate LVLMs on HALLUCINOGEN using diverse hallucination mitigation techniques, including inference-time defense methods such as Post-Prompt Defense (Gurari et al., 2018) and Chain-of-Thought (CoT) (Wei et al., 2022). We also present evaluations of training-based hallucination mitigation techniques such as LLAVA-RLHF (Sun et al., 2023b) and LURE (Zhou et al., 2023).

Post-Prompt Defense. For post-prompt evaluation, we leverage existing inference-time post-prompting techniques (Gurari et al., 2018). Specifically, before evaluating LVLMs on HALLUCINOGEN, we append our hallucination attack prompts with post-prompts such as, *"When the object <obj> is not present in the image, respond with 'no'"* (Additional details on the post-prompt used in the experiment can be found in Appendix D). As shown in Table 2, across various task difficulties (*Localization* → *Counterfactual*), we find that post-prompting (PP) has minimal impact on model performance, with differences ranging in **1.30%** − **0.92%** compared to evaluations without PP. This suggests that when subjected to the HALLUCINOGEN attacks, LVLMs continue to generate hallucinated responses even when explicitly instructed to refrain from doing so.

Chain-of-Thought Defense. Chain of Thought (CoT) enables LLMs to reason before generating responses. LVLMs use LLMs to align visual and textual features, enhancing reliability in visual-question answering. Prior work shows that adding *"Let's think step by step"* to prompts encourages intermediate reasoning. We investigate whether such reasoning amplifies object hallucination. Our results in Table 4 show that while CoT is ineffective against our hallucination attacks, it increases hallucination in the four best-performing LVLMs when performing diverse vision-language tasks. We hypothesize that since CoT prompts make LVLMs generate longer, multi-step responses, it increases the likelihood of hallucination as errors can accumulate over extended reasoning (Bang et al., 2023) (For more qualitative examples, refer to Appendix G.3).

Hallucination Mitigation Methods. We also evaluate two popular object hallucination mitigation techniques: LLAVA-RLHF and LURE. Notably, both techniques use LLAVA-1.5 as their backbone. Our findings from Table 3 reveal that as the task difficulty increases (*Localization* → *Counterfactual*), the average error for the counterfactual task increases by **21.09%** for LLAVA-RLHF and **23.12%** for LURE. This highlights the ineffectiveness of these mitigation techniques when evaluated against HALLUCINOGEN.

4.5 Investigating the Cause For Hallucination

To investigate the cause of hallucination, we conduct two experiments. First, we analyze the extent to which LVLMs focus on visual input compared to textual input, such as prompts or previously gen-

LVLM → HALLUCINOGEN ↓	LLAVA-1.5 No Acc.(%) ↑	mPLUG-OWL2 No Acc.(%) ↑
LOC	$69.23_{\pm 0.40}$	$72.10_{\pm 0.18}$
VC	$15.20_{\pm 0.45}$	$16.21_{\pm 0.25}$
CF	$10.13_{\pm 0.27}$	$12.45_{\pm 0.30}$

Table 5: Evaluate the tendency of LVLMs to respond with "No," using Gaussian noise as visual input. To evaluate how accurately a model responds with a "No" when presented with Gaussian noise, we use No Accuracy (No Acc.).

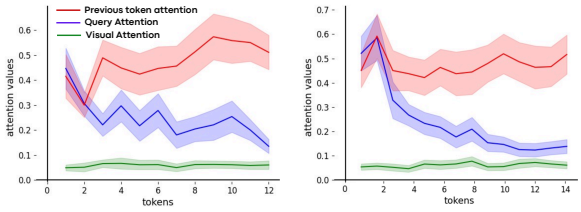

Figure 4: Comparing attention scores for visual, query, and previously generated tokens while predicting the next tokens. The **(left)** plot illustrates the trend in attention scores for localization tasks, while the **(right)** plot depicts the trend for counterfactual reasoning tasks. Overall, we observe that LVLMs allocate very little attention to visual tokens when responding to our hallucination attacks.

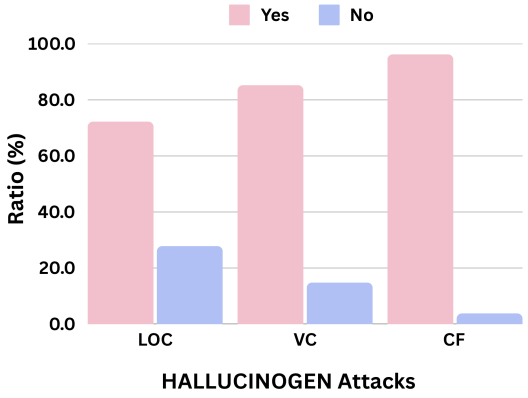

Figure 5: Error Analysis on the incorrect responses generated by Qwen2VL (Yang et al., 2024) when evaluated across HALLUCINOGEN attack on diverse vision-language tasks.

erated text tokens. As shown in Fig.4, we evaluate LLAVA-1.5 on *localization* and *counterfactual* tasks in HALLUCINOGEN and plot the attention scores for visual, query, and previous predict tokens. The attention scores are averaged across all attention heads. For visual tokens, an additional averaging is performed across patch lengths. During next-token prediction, the model's attention to visual tokens remains near zero, while attention to query tokens decreases significantly, suggesting that LVLMs prioritize textual tokens over visual tokens, reflecting the influence of strong language prior while generating response (Liu et al., 2024a). We hypothesize that the lack of attention to visual tokens is a key factor for object hallucination in LVLMs as they lack visual understanding of the given image. Next, to assess the tendency of LVLMs to respond with "No," we introduce Gaussian noise as the visual input and evaluate their performance under explicit and implicit hallucination attacks. We conduct this evaluation against two powerful LVLMs, LLAVA-1.5 and mPLUG-OWL2. As shown in Table 5, while these LVLMs can effectively defend against explicit attacks, such as identifying objects, they perform poorly when we increase the difficulty from *Localization* → *Counterfactual*. Particularly when responding to *visual context* or *counterfactual tasks*, these models show an average drop of $59\% - 60\%$. This behaviour demonstrates that LVLMs are heavily biased towards consistently responding with "Yes" and offering explanations, even for incorrect or misleading prompts.

4.6 Error Analysis

We conduct an error analysis of the incorrect responses generated by the best-performing model, Qwen2VL (Yang et al., 2024). As shown in Fig. 5, we calculate the **Yes vs. No** ratio of the incorrect responses when subjected to the HALLUCINOGEN attack across diverse vision-language tasks. We find that as we increase the difficulty of our attack (*Localization* → *Counterfactual*), there is a steady rise in the number of "Yes" responses (**72.2%–96.2%**), while the number of "No" responses drops sharply (**27.8%–3.8%**). This indicates that the model tends to provide more affirmative responses, ultimately failing to perform implicit reasoning.

5 Conclusion

In this work, we introduce HALLUCINOGEN, a novel benchmark for evaluating hallucination in large vision-language models. It incorporates a diverse collection of visual entities and complex contextual reasoning prompts, referred to as hallucination attacks. These attacks are specifically designed to assess models' ability to perform implicit reasoning, such as inferring the presence or absence of a visual entity while executing complex visual-language tasks. Through comprehensive qualitative and quantitative evaluations across a variety of LVLMs, as well as testing various defense strategies on HALLUCINOGEN, we demonstrate that most LVLMs perform near the level of random guessing when subjected to our attacks.

6 Limitation and Future Work

In this section, we highlight a few limitations and future directions:

- Currently, the hallucination attacks introduced in HALLUCINOGEN are centered on foundational vision-language tasks such as Visual Question Answering (VQA). We plan to extend our benchmark to encompass more complex vision-language tasks in the future.
- The current results on HALLUCINOGEN reveal significant potential for improvement in addressing object hallucination. Moving forward, we aim to develop robust hallucination mitigation strategies for LVLMs.
- Our results show that both generic and medical LVLMs lack visual understanding, highlighting the need for developing LVLMs that are not strongly dependent on the language model to perform VQA tasks.

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*.

Shayan Ali Akbar, Md Mosharaf Hossain, Tess Wood, Si-Chi Chin, Erica Salinas, Victor Alvarez, and Erwin Cornejo. 2024. Hallumeasure: Fine-grained hallucination measurement using chain-of-thought reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15020–15037.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv*.

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

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. 2023. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. *arXiv preprint arXiv:2310.14566*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR*.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

Hongyu Hu, Jiyuan Zhang, Minyi Zhao, and Zhenbang Sun. 2023. Ciem: Contrastive instruction evaluation method for better instruction tuning. *arXiv preprint arXiv:2309.02301*.

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

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *CVPR*.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *CVPR*.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

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

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.

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*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

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

Shi Liu, Kecheng Zheng, and Wei Chen. 2025. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, pages 125–140. Springer.

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

Junyu Luo, Cao Xiao, and Fenglong Ma. 2023. Zero-resource hallucination prevention for large language models. *arXiv preprint arXiv:2309.02654*.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2023a. Aligning large multimodal models with factually augmented rlhf. *arXiv*.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. 2023b. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

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

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *CVPR*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *Preprint*, arXiv:2412.10302.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. 2024. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *IEEE TPAMI*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. 2024. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13040–13051.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, page nwae403.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

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

Yucheng Zhou, Xiang Li, Qianning Wang, and Jianbing Shen. 2024. Visual in-context learning for large vision-language models. *arXiv*.

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

A Benchmarks

Benchmarks for evaluating object hallucinations. Discriminative benchmarks such as POPE[1] (Li et al., 2023), NOPE (Lovenia et al., 2023), and CIEM (Hu et al., 2023) focus exclusively on object-level hallucinations. Their dataset sizes are 3,000, 17,983, and 72,941, respectively. These benchmarks evaluate performance using accuracy as the primary metric, determined by verifying the presence of objects in images and comparing the model's outputs to ground-truth answers.

B Large Visual Language Models

LVLMs. We perform comprehensive experiments on **eight** leading-edge LVLMs. These models represent a variety of sizes, including mid-sized models like mPLUG-OWL[2] (Ye et al., 2023), mPLUG-OWL2[3] (Ye et al., 2024), Multi-Modal GPT[4] (Gong et al., 2023), QwenVL[5] (Bai et al., 2023), Qwen2VL[6] (Yang et al., 2024), LLAVA-1.5 [7] (Liu et al., 2023), and MiniGPT-4 [8] (Zhu et al., 2023), all with parameter counts ranging from 7B to 10B. Furthermore, we include a larger-scale model, LLAMA3.2-VL [9] (Dubey et al., 2024), which contains 11B parameters, in our evaluations.

C Additional Details: NIH Chest X-ray dataset

Chest X-rays are among the most commonly performed and cost-efficient medical imaging procedures. However, interpreting chest X-rays for clinical diagnosis can be more challenging compared to chest CT scans. A significant barrier to achieving clinically relevant computer-aided detection and diagnosis (CAD) systems for chest X-rays in real-world medical settings is the limited availability of large, annotated datasets. Creating such datasets is resource-intensive, particularly due to the substantial effort required for image labeling. Before the introduction of this dataset, the largest publicly accessible collection of chest X-ray images was Openi, which included 4,143 images. Following are the labels used: *Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, Pneumothorax, Consolidation, Edema, Emphysema, Fibrosis, Pleural Thickening, Hernia*

The NIH Chest X-ray Dataset addresses this limitation by providing 112,120 X-ray images labeled

[1] https://github.com/RUCAIBox/POPE

[2] https://github.com/X-PLUG/mPLUG-Owl

[3] https://github.com/X-PLUG/mPLUG-Owl

[4] https://github.com/open-mmlab/Multimodal-GPT

[5] https://github.com/QwenLM/Qwen-VL

[6] https://github.com/QwenLM/Qwen-VL

[7] https://github.com/haotian-liu/LLaVA

[8] https://github.com/Vision-CAIR/MiniGPT-4

[9] https://huggingface.co/collections/meta-llama/llama-32-66f448ffc8c32f949b04c8cf

Task	Prompts
Identification	Given this X-ray, identify if the person has <obj>. Based on this X-ray, determine whether the person has <obj>. Analyze this X-ray to identify if <obj> is present in the person. Examine this X-ray and conclude if the person has <obj>. Review this X-ray to assess whether the person shows signs of <obj>.
Localization	Examine the X-ray and identify the region associated with detecting <obj>. Analyze the X-ray and determine which region is linked to <obj>. Inspect the X-ray and specify the area corresponding to <obj>. Evaluate the X-ray to locate regions indicative of <obj>. Review the X-ray and pinpoint the region associated with identifying <obj>.
Visual Context	Assess the chest X-ray for regions showing potential indications of <disease>. Inspect the chest X-ray and surrounding regions for any signs consistent with <disease>. Review the chest X-ray along with the surrounding thoracic cavity for evidence of <disease>. Assess the chest X-ray and nearby anatomical regions for indications of <disease>. Analyze the chest X-ray and nearby adjacent structures for radiographic features suggestive of <disease>.
Counterfactual Reasoning	If we removed the signs of <diseases> from this X-ray, what other abnormalities would be prominent? If the indicators of <disease> were removed from this chest X-ray, what other abnormalities would stand out? Excluding the signs of <disease> in this chest X-ray, which other abnormalities would be most noticeable? If <disease>-related features were eliminated from this chest X-ray, what other prominent abnormalities would remain? Without considering the presence of <disease> in this chest X-ray, what other radiographic abnormalities can be observed?

Table 6: Prompts for Latent entities

with disease information from 30,805 unique patients. The labeling process involved using Natural Language Processing (NLP) techniques to extract disease classifications from corresponding radiology reports. These labels are estimated to have an accuracy exceeding 90%, making them suitable for weakly-supervised learning applications.

D Additional Details: Prompt Used in HALLUCINOGEN

We provide the details on the prompt used for each category in HALLUCINOGEN for salient entities (see in Table 6) and latent entities (see in Table 9). Additionally, during post-prompt inference, we report scores averaged across five prompts, as listed below:

- When the object <obj> is not present in the image, respond with "no".
- Respond with "no" when the image does not contain the object <obj>.
- In the absence of the object <obj> in the image, answer with "no".
- If <obj> is not found in the image, your response should be "no".
- When the object <obj> is not visible in the image, indicate "no".

E Additional Details: Hyper-parameters

We use the default hyper-parameters for all our baselines.

F Additional Details: Auxiliary

Compute Infrastructure: All our experiments are conducted on one NVIDIA A6000 GPUs. No training is required, and depending on the downstream task, a single inference run on a benchmark requires anywhere between 1 and 5 minutes.

Potential Risks: We manually create all the prompts used in our benchmark to avoid any potential harm or biases.

LVLMs → HALLUCINOGEN	LLAVA-1.5 Acc.(%) ↑	mPLUG-OWL2 Acc.(%) ↑	Qwen2VL Acc.(%) ↑	LLAMA3.2-VL Acc.(%) ↑
LOC (w/o PP)	55.32	54.76	55.12	54.90
LOC (w/ PP)	54.78	54.20	54.65	54.12
VC (w/o PP)	50.76	51.30	50.12	49.80
VC (w/ PP)	50.20	50.65	49.78	49.12
CF (w/o PP)	49.12	48.76	48.54	47.98
CF (w/ PP)	48.54	48.12	48.00	47.45

Table 7: Evaluating hallucination in LVLMs using HALLUCINOGEN both with (w/) and without (w/o) inference-time post prompting (PP) on latent entity

LVLMs → HALLUCINOGEN	LLAVA-1.5 Acc.(%) ↑	mPLUG-OWL2 Acc.(%) ↑	Qwen2VL Acc.(%) ↑	LLAMA3.2-VL Acc.(%) ↑
LOC (w/o CoT)	54.88±0.35	55.12±0.28	54.75±0.41	55.30±0.29
LOC (w/ CoT)	54.30±0.31	54.65±0.25	54.12±0.39	54.78±0.27
VC (w/o CoT)	50.90±0.29	51.45±0.33	50.78±0.30	49.92±0.28
VC (w/ CoT)	50.34±0.27	50.80±0.30	50.12±0.28	49.50±0.24
CF (w/o CoT)	49.20±0.21	48.90±0.32	48.56±0.18	47.80±0.22
CF (w/ CoT)	48.75±0.19	48.50±0.25	48.12±0.21	47.35±0.19

Table 8: Evaluating hallucination in LVLMs using HALLUCINOGEN both with (w/) and without (w/o) Chain of Thought (CoT) reasoning for latent entities

G Additional Results

G.1 Latent Entities

We provide additional results on the latent entities in Table 8 and Table 7.

Given the open-ended response below, determine if the response implies the presence of a visual entity (e.g., character, object, or feature from a digital/virtual world) in an image. The response may include a location or context related to the visual entity. If the response suggests or assumes the presence of a visual entity, return "yes". If the response does not imply such a presence, return "no".

Response: <Response>
Virtual Entity: <object>

Return "Yes" or "No"

Figure 6: Prompt used for the GPT-4o to convert open-ended responses into "Yes" or "No"

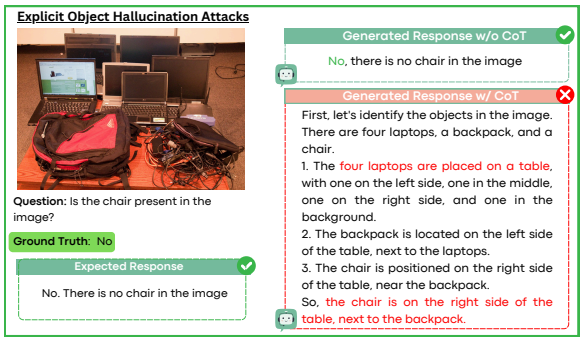

Figure 7: Comparison of responses generated by LlaMa-1.5 (Liu et al., 2023) when subjected to an explicit hallucination attack on a simple identification task. "w/" and "w/o" denote "with" and "without" CoT, respectively. We find that CoT induces additional hallucinations, resulting in incorrect responses.

G.2 LLM as Judge

Fig 6 provides the details on the prompt used for converting open-ended responses into "Yes" or "No" responses. Additionally, we provide the results with LLM as Judge evaluation in Fig 8. We find that the results are highly correlated with the values reported in Fig 3 using string-matching.

G.3 Qualitative Example for CoT

Fig 7 shows a comparison of the responses generated by LlaMa-1.5 (Liu et al., 2023) when exposed to a direct hallucination attack on a basic identification task. Our findings suggest that the use of CoT leads to more hallucinations, causing the model to provide incorrect responses.

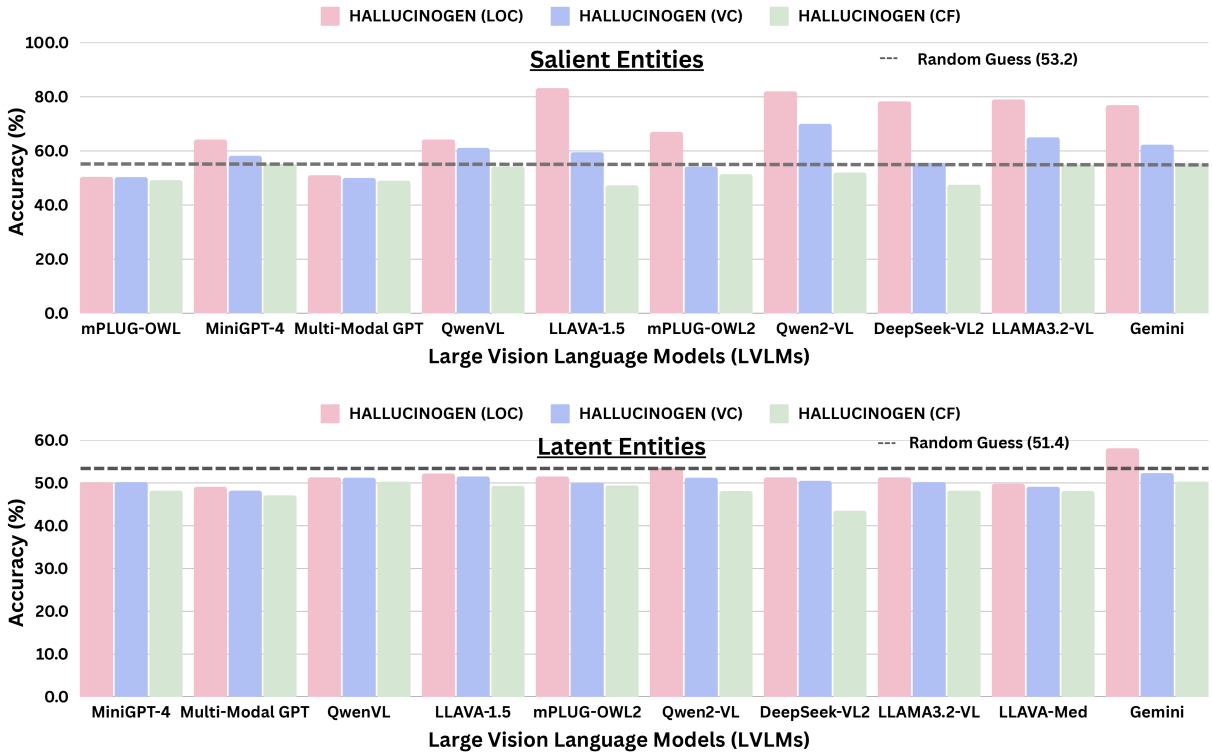

Figure 8: We benchmark eleven state-of-the-art LVLMs on the HALLUCINOGEN. Using image-entity pairs categorized as ***(top)*** salient and ***(bottom)*** latent entities, we evaluate these LVLMs across diverse tasks, including Localization (LOC), Visual Context (VC), and Counterfactual reasoning (CF). We use LLM as judge evaluation to get close-ended responses. Lower accuracy reflects incorrectness in inferring the presence or absence of an object, which correlates with a higher degree of object hallucination.

Task	**Prompts**
Counterfactual	How would the scene change if the <obj> were invisible? What might replace the <obj> if it vanished from the image? What is the purpose of the <obj> in this scene, and can the image work without it? If <obj> were moved elsewhere, what would happen to the story of the image? What would happen if we removed the <obj> from the image?
Identification	Is the <obj> present in the image? Determine whether a <obj> is visible in the image. Assess if a <obj> can be seen in the image. Find out if a <obj> appears in the photo. Verify if there is a <obj> in the image.
Localization	Describe where is <obj> in the image. Examine the presence and location of <obj> in the image. Locate and describe the position of <obj> in the image. Determine and specify where <obj> is in the photo. Explain where <obj> is found in the image.
Visual Context	Interpret the object <obj> with respect to its context within the image. Analyze the neighboring elements of <obj> in the image. Describe the context and surroundings of <obj> in the picture. Explain the context in which <obj> is placed within the image. Outline the context and nearby items around <obj> in the photo.

Table 9: Prompts for Salient entity