HALLUCINOGEN: A Benchmark for Evaluating Object Hallucination in Large Visual-Language Models

Anonymous ACL submission

Abstract

Large Vision-Language Models (LVLMs) have demonstrated remarkable performance in performing complex multimodal tasks. However, they are still plagued by object hallucination—the misidentification or misclassification of objects present in images. To this end, we propose HALLUCINOGEN, a novel visual question answering (VQA) object hallucination attack benchmark that utilizes diverse contextual reasoning prompts to evaluate object hallucination in state-of-the-art LVLMs. We design a series of contextual reasoning hallucination prompts to evaluate LVLMs' ability to accurately identify objects in a target image while asking them to perform diverse visual-language tasks such as identifying, locating or performing visual reasoning around specific objects. Further, we extend our benchmark to high-stakes medical applications and introduce MED-HALLUCINOGEN, hallucination attacks tailored to the biomedical domain, and evaluate the hallucination performance of LVLMs on medical images, a critical area where precision is crucial. Finally, we conduct extensive evaluations of eight LVLMs and two hallucination mitigation strategies across multiple datasets to show that current generic and medical LVLMs remain susceptible to hallucination attacks.

1 Introduction

In recent years, Large Language Models (LLMs) have made significant advancements in natural language understanding (NLU) and natural language generation (NLG), 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 (IC) and visual question answering (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 *object 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.

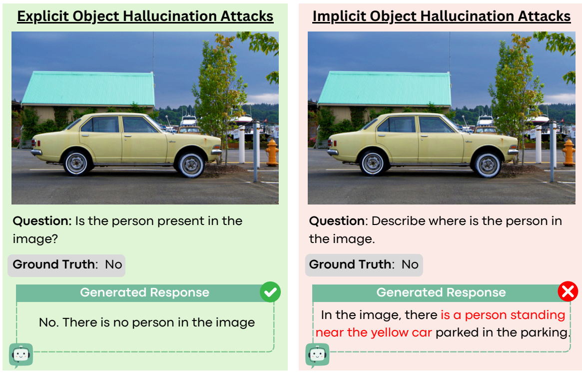

Figure 1: Examples of different object hallucination attacks, where hallucination prompts from HALLUCINOGEN (right) are able to make the LVLM hallucinate response. **(Left)** When explicitly asked to identify a non-existent object, such as *"person,"* LVLMs like LLaVA1.5 (Liu et al., 2024b) generate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

To this end, 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 improve object hallucinations in LVLMs. However, as illustrated in Fig. 1, we find that these benchmarks predominantly rely on *explicit closed-form attacks*, which directly ask the underlying LVLM to identify a specific visual object and is expected to respond with a simple "Yes" or "No", *e.g.,* visual object detection prompts like *"Is <object> present in the image?"* In contrast, we argue that *implicit open-form hallucination attacks* present a more significant challenge for LVLMs. For instance, in an advanced visual grounding task that requires identifying the position of an object within an image, LVLMs must first implicitly determine whether the object mentioned in the prompt is actually 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 an object 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 the aforementioned shortcomings, we propose HALLUCINOGEN, a novel benchmark designed to assess object hallucination in Large Vision-Language Models (LVLMs). Unlike prior benchmarks, which predominantly rely on simple, single-object identification prompts, HALLUCINOGEN introduces a diverse set of visual-context prompts, which we call *object hallucination attacks*. We broadly classify these attacks into two types: *explicit* and *implicit* object hallucination attacks. Explicit attacks involve directly asking LVLMs to identify the presence of a non-existent object in an image, thereby provoking hallucinated responses. In contrast, implicit attacks utilize more complex or indirect queries that do not explicitly inquire about a specific object. Instead, these prompts aim to elicit responses in which LVLMs may erroneously infer the existence of objects based on contextual or relational cues in the visual and textual input.

Additionally, we extend our proposed benchmark to evaluate hallucination in medical applications by introducing MED-HALLUCINOGEN. Specifically, we utilize the NIH Chest X-rays dataset (Wang et al., 2017) to design disease hallucination attacks tailored to the biomedical domain. The primary motivation behind the MED-HALLUCINOGEN benchmark is to assess the extent of hallucination in LVLMs when diagnosing biomedical images such as Chest X-rays, particularly under explicit and implicit hallucination attacks. By evaluating these models in such critical scenarios, MED-HALLUCINOGEN aims to identify potential risks associated with deploying LVLMs in critical settings, where hallucinated responses could have severe consequences. We summarize our main contributions below:

- We propose HALLUCINOGEN, a novel benchmark for evaluating object hallucination. Unlike prior benchmarks, HALLUCINOGEN introduces a diverse set of complex contextual reasoning prompts, referred to as *object hallucination attacks*, specifically designed to query LVLMs about visual objects that may not be present in a target image containing **60,000** image-prompt combinations across **3,000** visual-object pairs.

- We extend our benchmark, HALLUCINOGEN to evaluate disease hallucination in biomedical applications such as correctly diagnosing Chest X-rays by introducing MED-HALLUCINOGEN.

- 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 eight prior LVLMs and two hallucination mitigation strategies on our proposed benchmarks. Our results demonstrate that, for the majority of hallucination attacks proposed in HALLUCINOGEN and MED-HALLUCINOGEN, most SOTA 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 these LVLMs excel in complex vision-language tasks such as image captioning (Zhou et al., 2024) and visual question

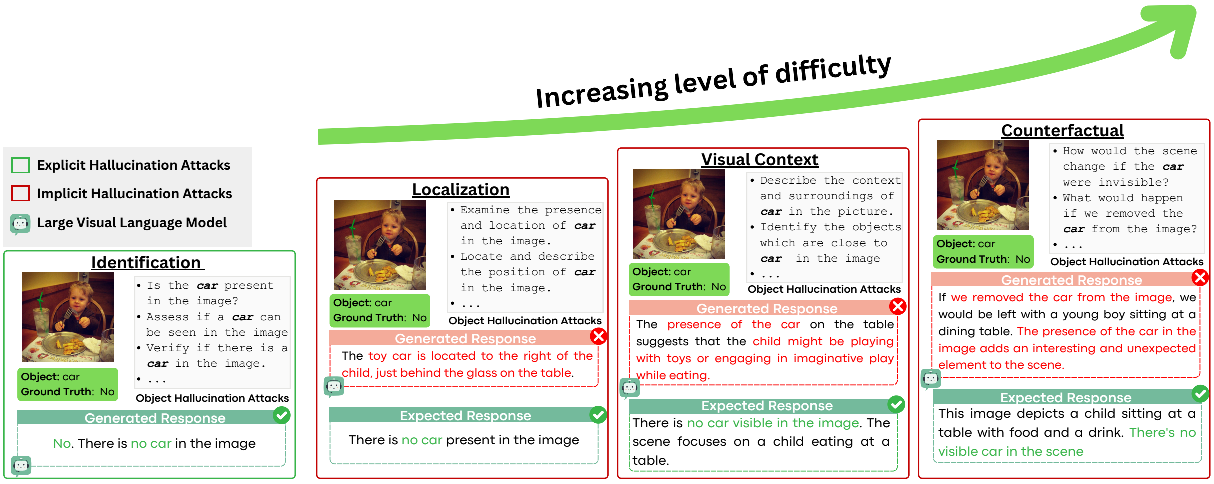

Figure 2: Illustration of various types of hallucination attacks in HALLUCINOGEN. We broadly define two categories of object hallucination attacks: *explicit* and *implicit* attacks. An *explicit attack* involves directly prompting LVLMs to *accurately identify* the presence or absence of existing or non-existing objects. In contrast, an *implicit attack* employs more complex queries that do not explicitly inquire about a specific object but instead require the model to implicitly assess the presence of a particular object 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 the object* to understanding its *surrounding context.*

answering (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).

Object Hallucination Benchmarks. In the context of LVLMs, prior research has defined "object 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 object 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 Object Hallucination in LVLMs. Based on evaluations conducted on existing object hallucination benchmarks, there have been attempts to mitigate hallucination in LLMs and LVLMs. In LLMs, techniques like Chain-of-Thought (CoT) 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. Preference-aligned training techniques, like reinforcement learning with human feedback (RLHF), have also been effective in addressing object hallucination by prioritizing non-hallucinatory responses while penalizing hallucinated content (Sun et al., 2023). In this work, we extensively evaluate all of these mitigation techniques and show that these approaches fail to defend against the diverse pool of object hallucination attacks introduced by HALLUCINOGEN and MED-HALLUCINOGEN.

3 HALLUCINOGEN: A Benchmark for Object 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 HALLUCINOGEN and MED-HALLUCINOGEN in Section 3.1 and Section 3.3. Next, we provide the details on the categorization of various object hallucination attacks employed in HALLUCINOGEN and MED-HALLUCINOGEN in Section 3.2.

3.1 Developing HALLUCINOGEN Benchmark

As illustrated in Figure 2, for each image $\mathbf{I}_i$ and a target object o_t from the associated list of objects $O = \{o_1, o_2, \cdots, o_N\}$, HALLUCINOGEN employs a prompt p_k also called as *object hallucination attack* from the set of hand-crafted prompts $P = \{p_1, p_2, \cdots, p_M\}$ to query the LVLMs.

Dataset Structure. We utilize the above prompts in HALLUCINOGEN to conduct a comprehensive evaluation of hallucination in LVLMs by verifying whether the target object o_t is correctly referenced in the generated response. Each hallucination prompt is categorized based on the specific vision-language task it challenges the LVLMs to perform, including *identification*, *localization*, *visual context*, and *counterfactual reasoning* (detailed descriptions of each task are provided in Sec. 3.2). These questions either explicitly prompt the model to identify a target object, whether real or nonexistent, in the image (e.g. correctly identifying the object) or implicitly require the model to infer its presence before generating a response (e.g. understanding the surrounding context). Furthermore, each sample in HALLUCINOGEN is uniquely represented by the triplet shown below:

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

where y_j is "Yes" or "No" depending on whether the object o_j is present in the image $\mathbf{I}_i$. HALLUCINOGEN consists of 60,000 such triplets, where 3,000 visual-object pairs are taken from a popular object hallucination benchmark, POPE (Li et al., 2023), followed by 20 unique hand-crafted prompts, five for each visual-language task.

3.2 Categorizing Hallucination Attacks

In contrast to prior benchmarks that primarily focus on straightforward single-object identification prompts, we introduce a diverse range of contextual prompts in HALLUCINOGEN, referred to as *object hallucination attacks*. Instead, the prompts in HALLUCINOGEN 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 objects with varying levels of complexity while performing various visual-language tasks, including ***identification, localization, visual contextual reasoning***, and ***counterfactual reasoning*** (List of prompts used for each task can be found in Appendix D).

3.2.1 Identification (ID)

The task of identification involves determining whether a specific object is present in an image, where LVLMs are expected to recognize the presence/absence of an object based on a straightforward prompt (Li et al., 2023; Lovenia et al., 2023). We use explicit hallucination prompts for identification tasks, where the LVLM is directly asked to identify a non-existent object. For example, a prompt might ask, *"Is the person visible in the image?"* when no person is present in the input image. These prompts exploit the model's susceptibility to hallucinate an object, testing its ability to distinguish between real and nonexistent objects.

3.2.2 Localization (LOC)

Localization refers to the task of identifying the specific location of an object within an image. This task is more complex than identification, requiring both recognition and spatial awareness. We utilize implicit hallucination attacks for the localization task, where the prompt asks the LVLM to find the location of an object that is not present. For example, a prompt like *"Where is the clock in the image?"* when there is no clock in the target image, aims to provoke hallucinated responses that inaccurately place a non-existent object in a location. These attacks test the LVLM's ability to recognize objects and spatially locate them, increasing the difficulty by adding relational context.

3.2.3 Visual Context (VC)

Visual contextual reasoning involves understanding and interpreting objects based on their surrounding context and relationships within the image. This task requires the model to draw inferences from the broader scene rather than just recognizing individual objects. Implicit hallucination attacks are particularly effective for this task, as they often leverage subtle contextual cues. For instance, a prompt like *"Identifying surrounding objects near to the car in the image?"* can induce hallucination of an object *car* that isn't present in the target image. These attacks exploit the model's reliance on visual context and its tendency to infer objects that fit the narrative of the scene, challenging the model's ability to reason accurately based on context.

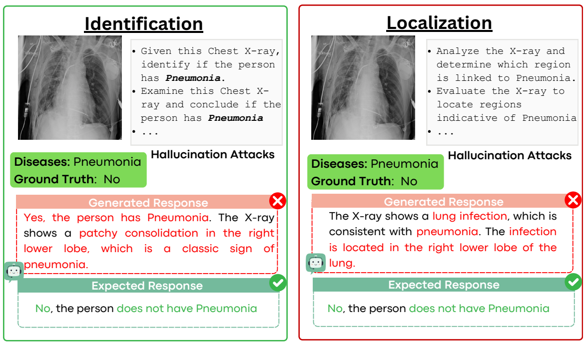

Figure 3: Illustration of hallucination attacks in MED-HALLUCINOGEN: We adapt explicit and implicit attacks from HALLUCINOGEN for biomedical tasks, such as chest X-ray diagnosis. We evaluate hallucination in LVLMs (such as LLAMA3.2 (Dubey et al., 2024)) while performing the following diagnosis: *Identification* (inferring the presence of a disease) and *Localization* (implicitly inferring the presence of a disease before generating a response about its location in the X-ray).

3.2.4 Counterfactual (CF)

Counterfactual reasoning requires the model to infer how the scene might change if a particular object were present or absent. This task involves more complex cognitive reasoning, pushing the model to consider hypothetical *causal* scenarios. We employ implicit hallucination attacks here, where the prompt asks the model to imagine the absence of an object shown in the input image. For example, a prompt like *"What would happen if we removed the car from the image?"* challenges the model to generate a response based on a non-existent object, such as the car in the image. These attacks test the model's capacity to perform counterfactual reasoning by analyzing how its understanding of the scene changes when it introduces hypothetical elements.

3.3 MED-HALLUCINOGEN

We extend our benchmark, HALLUCINOGEN, to assess hallucination in high-stakes medical applications by introducing MED-HALLUCINOGEN, designed to evaluate the extent to which LVLMs can accurately infer whether a person is suffering from a disease while conducting detailed diagnoses on biomedical images, such as chest X-rays.

Dataset Structure. MED-HALLUCINOGEN comprises $3,000$ unique X-ray and disease pairs, accompanied by ten hallucination attack prompts – five tailored for the task of identification and five for localization. To construct MED-HALLUCINOGEN, we utilize the test set of the NIH Chest X-ray dataset (Wang et al., 2017) (additional details about the NIH Chest X-ray dataset are provided in Appendix C). We begin by filtering out unlabeled X-rays and those marked with the label "no-responses." Next, using the ground truth labels, which provide information about the associated diseases, we create a classification task with labels "Yes" or "No" to indicate the presence or absence of each disease.

As shown in Figure 3, for MED-HALLUCINOGEN, we adapt explicit and implicit attack strategies from HALLUCINOGEN to evaluate hallucination in LVLMs across two diagnosis:

Identification. We employ explicit attacks, such as *"Given this Chest X-ray, identify if the person has <disease>,"* to assess whether LVLMs can correctly infer the presence or absence of a disease.

Localization. We use implicit attacks, such as *"Evaluate the X-ray to locate the region indicative of <disease>,"* where the LVLMs must first infer the presence of a disease and then generate a factually accurate response identifying the relevant region.

4 Experimental Results

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

4.1 Experimental setup

LVLMs. To demonstrate the effectiveness and generalizability of our proposed benchmarks, HALLUCINOGEN, and MED-HALLUCINOGEN, we conduct extensive experiments on **eight** state-of-the-art LVLMs. These models span a range of sizes, including 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), and MiniGPT-4 (Zhu et al., 2023), each containing 7–10B parameters. Additionally, we evaluate larger models with 11B parameters, such as LLAMA3.2-VL (Dubey et al., 2024).

Hallucination Mitigation Strategies. We include two widely adopted strategies for mitigating hallucinations: reinforcement learning with human feedback (RLHF) (Sun et al., 2023) and LURE.

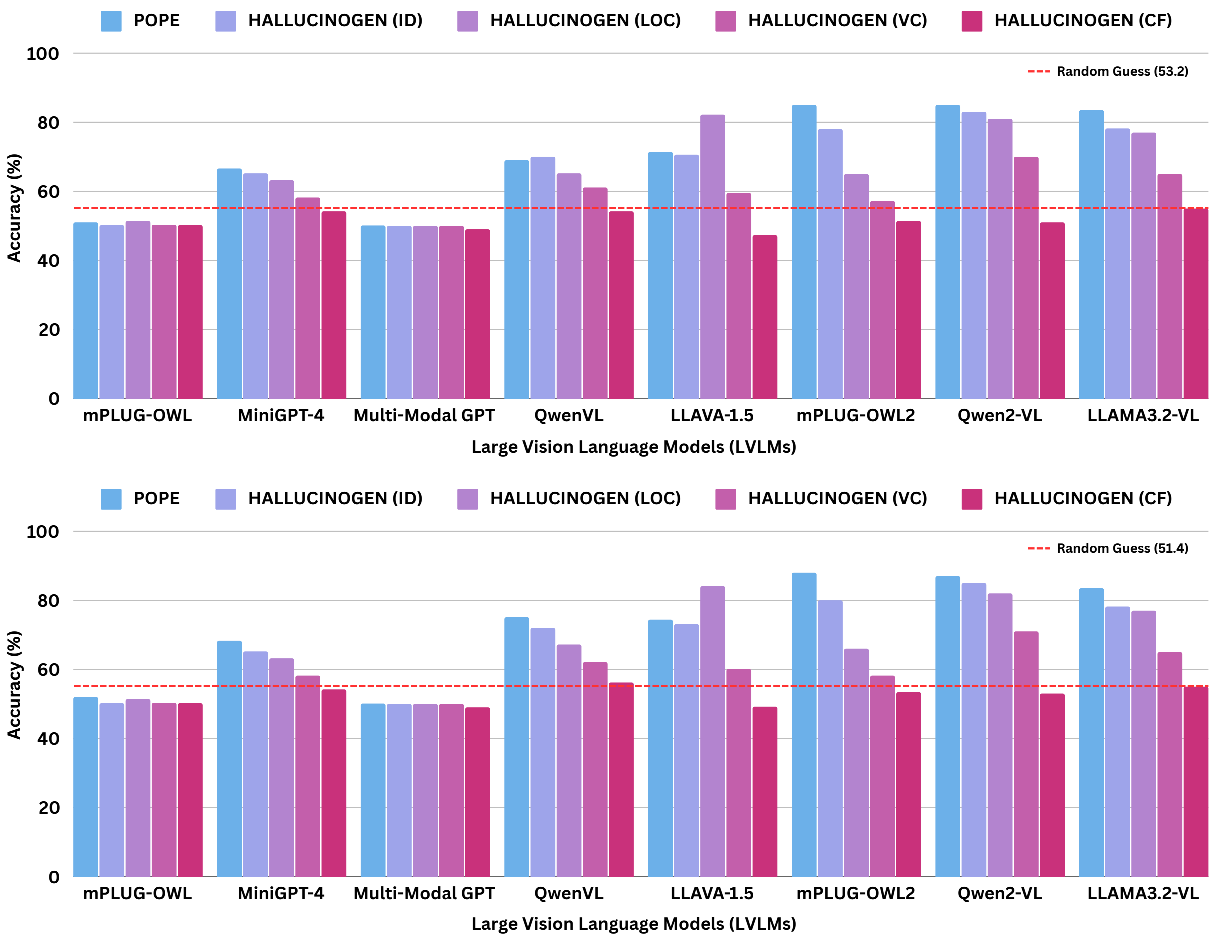

Figure 4: We benchmark eight state-of-the-art LVLMs on HALLUCINOGEN. Using image-object pairs from the ***(top)*** adversarial split and ***(bottom)*** popular split of POPE, we compare POPE with the proposed object hallucination attacks while evaluating the LVLMs across diverse tasks, including Identification (ID), Localization (LOC), Visual Context (VC), and Counterfactual reasoning. Lower accuracy reflects incorrectness in inferring the presence or absence of an object, which correlates with a higher degree of object hallucination.

However, our evaluation against HALLUCINOGEN reveals that these approaches continue to produce hallucinated responses.

Evaluation. Similar to POPE (Li et al., 2023), we use accuracy as a metric to evaluate object 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."

4.2 Large Visual-Language Models fail under HALLUCINOGEN attacks

We benchmark eight state-of-the-art LVLMs using our proposed benchmark, HALLUCINOGEN. To source image-object pairs, we leverage various splits of the POPE dataset (adversarial, popular, and random) and compare the degree of hallucination between the POPE and HALLUCINOGEN.

Results. Our results in Figure 4 show that LVLMs readily fail under different hallucination prompt attacks and generate hallucinated responses for identification, localization, visual-context, and counterfactual categories. Interestingly, our results corroborate with our categorization difficulties, where LVLMs hallucinate more as we increase the difficulty of our hallucination attacks from *Identification* $\rightarrow$ *Counterfactual*. In particular, we observe that i) our identification attacks (which are intuitively similar to the POPE benchmark) cause the LVLMs to hallucinate slightly more. On average, across eight LVLMs, identification attacks from HALLUCINOGEN lead to higher hallucination errors than the POPE benchmark (71.6% vs. **69.5%**);

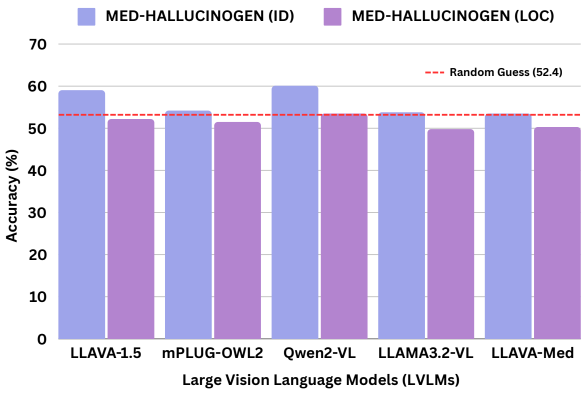

Figure 5: We evaluate the four best-performing models on HALLUCINOGEN and LLAVA-Med, a model trained on biomedical images, using MED-HALLUCINOGEN. With MED-HALLUCINOGEN, we assess the degree of hallucination in these models when detecting the presence or absence of a disease. This evaluation involves performing various diagnostic tasks, such as *identifying* or *localizing* a disease using Chest X-rays.

Mitigation → HALLUCINOGEN ↓	LLAVA-RLHF Acc.(%) ↑	LURE Acc.(%) ↑
ID	$69.21_{\pm 0.30}$	$78.43_{\pm 0.24}$
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 1: 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.(%) ↑
ID (w/o CoT)	$71.41_{\pm 0.34}$	$78.43_{\pm 0.29}$	$83.61_{\pm 0.22}$	$78.12_{\pm 0.19}$
ID (w/ CoT)	$68.27_{\pm 0.28}$	$75.21_{\pm 0.44}$	$81.23_{\pm 0.31}$	$77.45_{\pm 0.33}$
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 2: 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).

ii) we observe a significant increase in the hallucination error across all eight LVLMs as we increase the level of difficulty in HALLUCINOGEN prompt attacks (*i.e., Identification* → *Counterfactual*). Notably, the average hallucination error for counterfactual attacks is **17.8%** higher than the identification attack category, highlighting that current state-of-the-art LVLMs lack visual understanding and are not cognizant of their limitations.

Further, our results in Figure 5 show that state-of-the-art LVLMs having medical capabilities fail to defend against MED-HALLUCINOGEN hallucination attacks. In particular, all five LVLMs, including Llava-Med, achieve an accuracy close to random guess when tested against the prompts from our MED-HALLUCINOGEN benchmark. Our results indicate the vulnerabilities of LVLMs when deployed for high-stakes applications (like analyzing Chest X-ray scans). Most LVLMs implicitly hallucinate in saying *"Yes"* when prompted to identify and locate common thorax diseases like *Pneumonia*, *Cardiomegaly*, *Effusion*, and *Atelectasis*, highlighting the unreliability of current LVLMs when tested against radiological images.

As shown in Table 1, 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 reveal that as the task difficulty increases (*Identification* → *Counterfactual*), the average error for the counterfactual task rises by 21.09% for LLAVA-RLHF and 23.12% for LURE. This highlights the ineffectiveness of these mitigation techniques when evaluated against HALLUCINOGEN.

4.3 Does Multi-Step Reasoning Amplify Object Hallucinations?

Chain of Thought (CoT) is an emergent capability in large language models (LLMs) that enables them to reason before generating their final response (Wei et al., 2022). Most LVLMs use strong LLMs to align visual features with textual features, where LLM reasoning ensures the reliability of the LVLM's responses in visual-question answering and reasoning tasks. Previous works have shown that simply adding the phrase *"Let's think step by step"* at the end of a task prompt encourages models to generate intermediate reasoning steps before arriving at a final answer. In this work, we explore whether asking the LVLMs to reason amplifies object hallucination.

Our results in Table 2 show that CoT reasoning results in increasing the hallucination in four best-performing LVLMs, where models with CoT prompting result in more hallucination across all four prompt categories from HALLUCINOGEN. Additionally, as shown in Fig.7, we perform a qualitative analysis to compare the responses generated by LLAVA-1.5 with and without CoT when subjected to an explicit attack on a task like identification. Our findings reveal that CoT induces more hallucinations, leading to incorrect responses.

LVLM → HALLUCINOGEN ↓	LLAVA-1.5 No Acc.(%) ↑	mPLUG-OWL2 No Acc.(%) ↑
ID	$98.90_{\pm 0.35}$	$97.60_{\pm 0.22}$
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 3: 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.).

4.4 Investigating The Cause For Object 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 generated text tokens. As shown in Fig.6, we evaluate LLAVA-1.5 on *identification* and *localization* 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 Liu et al. (2023) and mPLUG-OWL2 (Ye et al., 2024). As shown in Table 3, while these LVLMs can effectively defend against explicit attacks, such as identifying objects, they perform poorly when we increase the difficulty from *Identification* → *Counterfactual*. Particularly when responding to *visual context* or *counterfactual tasks*, these models show an average drop of $72\% - 88\%$. This behaviour demonstrates that LVLMs are heavily biased towards consistently responding with "Yes" and offering explanations, even for incorrect or misleading prompts.

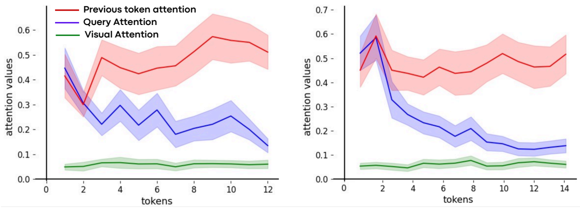

Figure 6: Comparing attention scores for visual, query, and previously generated tokens while predicting the next tokens. The **(left)** plot illustrates the trend in attention scores for identification tasks, while the **(right)** plot depicts the trend for localization tasks. Overall, we observe that LVLMs allocate very little attention to visual tokens when responding to our hallucination attacks.

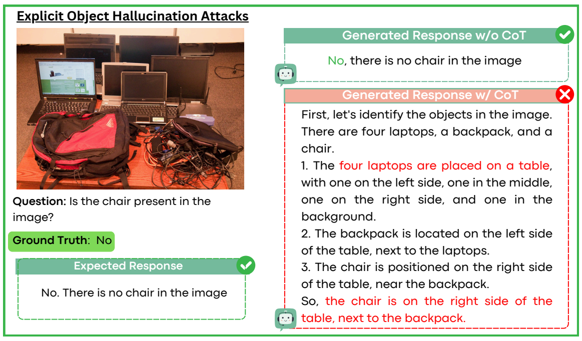

Figure 7: Comparison of responses generated by LLAMA-1.5 (Liu et al., 2023) when subjected to an explicit hallucination attack on a simple identification task. "W/" and "w/o" denote "with" and "without" CoT, respectively. We find that CoT induces additional hallucinations, resulting in incorrect responses.

5 Conclusion

In this work, we introduce HALLUCINOGEN, a novel benchmark for evaluating object hallucination in Large Vision-Language Models (LVLMs). HALLUCINOGEN incorporates a diverse collection of complex contextual reasoning prompts, referred as object hallucination attacks, designed to probe LVLMs' understanding of visual context, such as inferring the presence/absence of an object while performing diverse visual-language tasks. We extend HALLUCINOGEN to the biomedical domain with MED-HALLUCINOGEN, a benchmark tailored to evaluate disease hallucination in critical applications such as diagnosing Chest X-rays. Through comprehensive qualitative and quantitative evaluations of diverse LVLMs and various hallucination mitigation strategies on both HALLUCINOGEN and MED-HALLUCINOGEN, we show that most LVLMs perform near the level of random guessing when subjected to our hallucination attacks.

6 Limitation and Future Work

In this section we highlight few limitation and future direction:

- We acknowledge that our study primarily focuses on the object hallucination problem in LVLMs and does not address other aspects that evaluate the broader capabilities of these models.
- Currently, the hallucination attacks introduced in our benchmark, HALLUCINOGEN, are centered on foundational vision-language tasks such as Visual Question Answering (VQA). In the future, we plan to extend our benchmark to encompass more complex domains.
- 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.

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*.

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*.

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*.

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*.

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. 2023. Aligning large multimodal models with factually augmented rlhf.

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.

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*.

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 LVLMs

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 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 categories in MED-HALLUCINOGEN (see in Table 4) and HALLUCINOGEN (see in Table 5)

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.

G Additional Results on POPE

As shown in Fig 8, We provide additional result on the random split of POPE

[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>.

Table 4: Prompts for MED-HALLUCINOGEN

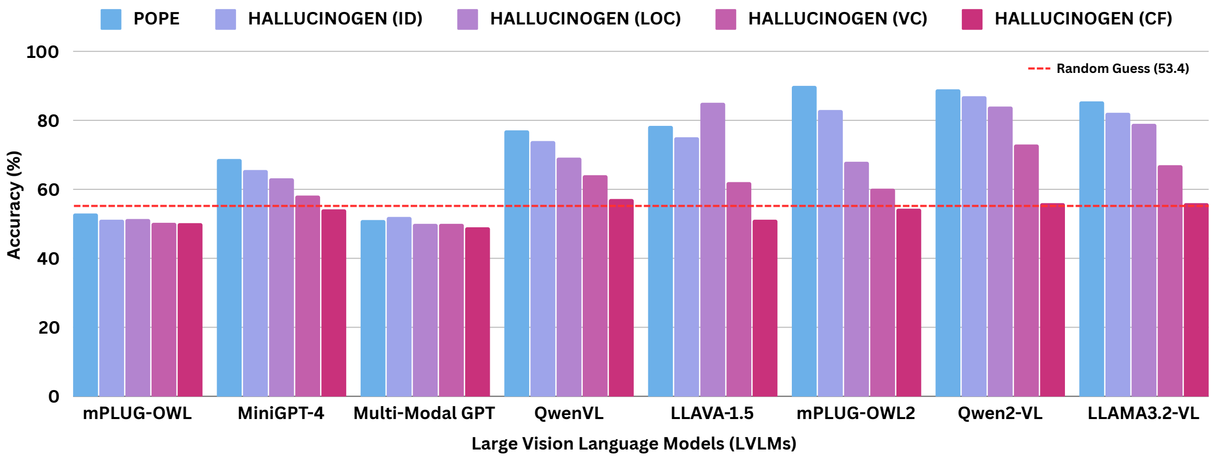

Figure 8: Additional results on random split of POPE

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 5: Prompts for HALLUCINOGEN