

Mirage in the Eyes: Hallucination Attack on Multi-modal Large Language Models with *Only* Attention Sink

Yining Wang¹, Mi Zhang^{1,*}, Junjie Sun¹, Chenyue Wang¹, Min Yang^{1,*},
Hui Xue², Jialing Tao², Ranjie Duan², Jiexi Liu²

¹ Fudan University, ² Alibaba Group

{ynwang22@m., mi_zhang@, jjsun22@m., wangcy23@m., m_yang@}fudan.edu.cn

{hui.xueh, jialing.tjl, ranjie.drj, liujiexi.ljx}@alibaba-inc.com

Abstract

Fusing visual understanding into language generation, Multi-modal Large Language Models (MLLMs) are revolutionizing visual-language applications. Yet, these models are often plagued by the *hallucination problem*, which involves generating inaccurate objects, attributes, and relationships that do not match the visual content. In this work, we delve into the internal attention mechanisms of MLLMs to reveal the underlying causes of hallucination, exposing the inherent vulnerabilities in the instruction-tuning process.

We propose a novel *hallucination attack* against MLLMs that exploits *attention sink* behaviors to trigger hallucinated content with minimal image-text relevance, posing a significant threat to critical downstream applications. Distinguished from previous adversarial methods that rely on fixed patterns, our approach generates dynamic, effective, and highly transferable visual adversarial inputs, without sacrificing the quality of model responses. Comprehensive experiments on 6 prominent MLLMs demonstrate the efficacy of our attack in compromising black-box MLLMs even with extensive mitigating mechanisms, as well as the promising results against cutting-edge commercial APIs, such as GPT-4o and Gemini 1.5. Our code is available at <https://huggingface.co/RachelHGF/Mirage-in-the-Eyes>.

1 Introduction

Integrating visual comprehension into language models, Multi-modal Large Language Models [16, 46] enable interaction with users across various modalities, and provide responses that demonstrate a deep understanding of complex visual semantics. Through instruction-tuning in multi-modal spaces, MLLMs have significantly advanced vision-language tasks like image captioning [6], visual grounding [58], and multi-modal conversations [49].

Despite their revolutionary impact, MLLMs face a significant challenge: *the hallucination problem*. This occurs when

they generate irrelevant or entirely fabricated responses according to the image content. Examples include mentioning non-existent objects [12, 38], providing inaccurate attributes [47], or describing inconsistent relationships between objects [68]. As MLLMs are increasingly employed to aid decision-making, task planning, and user interaction in critical fields like medical reasoning [36], autonomous driving [74], and robotic manipulation [37], their tendency to generate hallucinated responses poses significant risks. In applications with user-defined inputs, such as AI assistants [53], customer services [9], and physical therapy tools [84], manipulated contents may be injected or spread online, compromising system reliability and potentially misleading users without domain expertise.

To address the unintended effects of unfaithful MLLMs, the causes of hallucinations have been explored [4]. Some suggest that hallucinations arise from the imbalance between weak vision models and powerful LLM backbones [25, 35], which causes MLLMs to over-rely on language priors (e.g., the tendency of associating *bananas* with the color *yellow* more often than *green*). Other studies point to the statistical biases in MLLM pre-training datasets [35, 85], which often feature imbalanced object distributions and co-occurrence patterns, resulting in MLLMs generating irrelevant descriptions. While various studies have explored the causes of hallucination, they often focus on individual factors in isolation, largely in a post-hoc manner. In light of these gaps, our work explores the generation mechanisms of MLLMs, investigating the complex interactions between modalities to provide a comprehensive understanding of this phenomenon.

Recent research has identified a phenomenon known as *attention sink* [72, 80], where certain tokens receive extremely high attention scores during the generation of LLM responses. Further research on MLLMs extends this concept to multi-modal settings [28], showing that the sink token exhibits a unique columnar pattern within attention maps (as illustrated in Fig. 1), drawing significant attention in the subsequent generative processes. Notably, the hallucinated outputs are observed to generally follow these sink tokens, indicating a

*Corresponding authors.

potential connection between them. Through an in-depth examination of the instruction-tuning process in multi-modal training, we identify the critical flaw that MLLMs tend to produce irrelevant image-text content after following user instructions, while the aggregation of misleading global information further exacerbates the divergence from the actual image content.



Human: Please describe this image in detail.

Assistant: The image depicts a lively scene in a park, where two men are competing to catch a frisbee in mid-air...There are several other people in the park, some sitting on the grass and others standing around. In addition to the people, there are several balls scattered throughout the scene, possibly belonging to the park-goers. A car can also be seen parked in the background, adding to the overall atmosphere of the park.

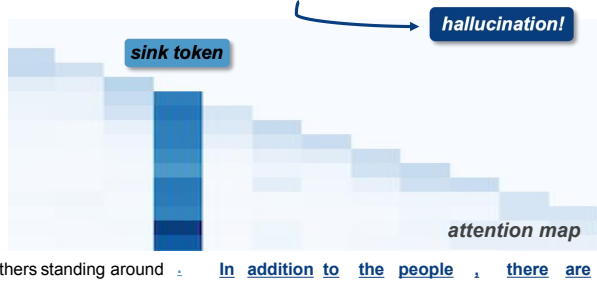


Figure 1: An illustration of the attention sink phenomenon in MLLM responses. The sink token receives high attention scores in a columnar pattern. The hallucinated responses are marked bold with **indigo tokens**.

Uncovering the formation of attention sink in hallucinated responses, we propose the first-ever *hallucination attack* against MLLMs with *only* attention sink. This pioneering attack is designed to exacerbate hallucinations in MLLM responses while maintaining their overall quality and utility. Our proposed attack overcomes the constraints of existing adversarial attacks against LLMs and MLLMs, which depend heavily on predefined target responses and task-specific datasets. By manipulating attention scores and hidden embeddings to induce sink tokens, our attack constructs dynamic, highly effective, and black-box transferable adversarial visual inputs. This method not only circumvents current mitigation strategies for hallucinations but also shows significant impacts on the latest commercial MLLM APIs. We hope this hallucination attack will expose the critical vulnerability of MLLMs in downstream applications, and drive advancements toward more reliable and high-performing multi-modal models.

Our Distinction from Previous Attacks. Adversarial attacks on MLLMs have been a concern since their development. These attacks aim to provoke harmful model re-

sponses [51, 59] or to impair model performance on specific tasks [15, 21], but they still rely heavily on predefined patterns or task-specific datasets. For instance, perturbation-based attacks [59, 82] optimize adversarial perturbations according to predefined target responses, which require massive human labor in crafting and filtering. Other text-based attacks insert human-written jailbreak templates [50, 51] (e.g., role-playing scenarios) to disrupt the safety alignments, but often lack transferability across different models. Recent structure-based attacks [13, 62] embed harmful instructions into images with typography or text-to-image models, which also require carefully designed textual triggers to redirect the model’s attention. Concentrating on adversarial visual inputs, our work is distinct from previous adversarial attacks in two key aspects:

1. **Attack Objective:** Our hallucination attack aims to induce the generation of erroneous objects, attributes, and relationships in MLLM responses, whereas previous adversarial attacks primarily focus on triggering harmful outputs and bypassing safety alignments.
2. **Attack Efficiency:** We achieve a dynamic and effective attack by directly manipulating the self-attention mechanisms during the MLLM generation process. This approach overcomes the limitations of previous methods, which demand substantial human efforts to define the target behaviors of models.

Our Contributions are summarized as follows.

- We present a comprehensive analysis linking the attention sink phenomenon with hallucination issues in MLLMs. By probing the inherent limitations of the instruction-tuning process, we expose the model’s tendency to produce two-segment responses with declined image-text relevance, where the aggregation of misleading information contributes to the hallucinated outputs.
- We propose the first hallucination attack targeting MLLMs with only the manipulation of attention sinks. The crafted adversarial visual inputs significantly exacerbate the object, attribute, and relationship hallucination without degrading response quality. By manipulating the attention mechanism and hidden embeddings, our attack achieves high transferability and adaptability without relying on predefined patterns.
- The extensive evaluation, assisted by GPT-4 [1], assesses our attack in hallucination and response quality on six prominent MLLMs. Remarkable results demonstrate that our attack transfers effectively to black-box MLLMs and commercial APIs such as GPT-4o [57] and Gemini 1.5 [56], successfully overcoming three categories of existing mitigation strategies. Up to 10.90% and 12.74% increase in hallucinated sentences and words highlight the vulnerability of critical downstream applications to our proposed attack.

2 Related Work

2.1 Multi-modal Large Language Models

The surge of LLMs has highlighted their remarkable capabilities in in-context learning, instruction following, and multi-step reasoning. MLLMs build on these strengths by incorporating additional modalities such as image, video, and audio, enabling users to engage with both textual and multi-modal prompts. MLLMs typically consist of three key components: multi-modal encoders (e.g., Vision Transformers [18] as visual encoders), pre-trained LLM backbones, and the adapter modules for modality alignment. The forerunner MLLMs like MM-GPT [23] achieve cross-modality alignment by integrating gated cross-attention blocks within their LLM backbones, but are constrained by massive computational demands [5]. On the other hand, models like LLaVA [46] and Shikra [10] utilize linear projection layers to map multi-modal features into textual spaces, offering a more computationally efficient solution. Moreover, advanced MLLMs such as InstructBLIP [16] and mPLUG-owl [76] incorporate Q-former modules, which consist of two Transformer blocks with shared self-attention layers. These modules update a set of learnable queries to effectively integrate both textual and multi-modal features. Most MLLMs follow a two-stage training paradigm: first pre-training on large-scale datasets to bridge modality gaps, and then instruction-tuning on task-related data to enhance multimodal conversational capabilities.

2.2 Mitigating Hallucination in MLLMs

Recent efforts to mitigate the hallucination problem in MLLMs fall into three categories: mitigation through *decoding*, *model retraining*, and *post-processing*.

Some work optimizes the decoding strategy during the inference stage, to suppress the generation of hallucinated responses. For instance, OPERA [28] mitigates hallucination by penalizing columnar attention patterns during beam search decoding, thereby reducing the model’s over-reliance on certain summary tokens. Meanwhile, VCD [35] calibrates model output distributions with both clean and distorted inputs, effectively counteracting the language priors of LLM backbones. In the realm of model retraining, HACl [29] employs contrastive learning of multi-modal representations to distinguish between hallucinated and factual responses, while [81] adjusts the prediction of EOS token to prevent excessively long hallucinated texts. Other retraining-based methods gather high-quality visual instruction tuning data to correct statistical biases [78], or resort to human feedback as a reward for response correction [79]. Another category of work leverages post-hoc techniques to rectify hallucinations, such as training additional hallucination revisors [34], or implementing multi-agent debate to examine the faithfulness of model responses [42].

2.3 Attention Sink

Transformer-based models leverage the self-attention mechanism to capture global information [67]. However, during inference, it has been observed that some tokens receive disproportionately high attention compared to others, which is known as the *attention sink* phenomenon.

The phenomenon was first observed in Large Language Models (LLMs) [72], where initial tokens, despite their low semantic significance, receive high attention. One reasonable explanation focuses on the sum-to-one property of the Softmax operation within the self-attention mechanism, where redundant attention is allocated to the initial tokens, as they are always visible during the generation process. The phenomenon has spurred several innovative approaches in LLM quantization [63] and LLM attention calibration [80].

When the attention mechanism is applied across different modalities, the attention sink is observed to cause hallucination responses [28]. This occurs when certain sink tokens, characterized by columnar patterns with high attention scores, dominate the prediction of subsequent tokens, overshadowing all preceding responses (as depicted in Fig. 1). As a result, the models will neglect image tokens and prior context, ultimately resulting in hallucination responses.

3 Attention Sink and Hallucination

In the following section, we analyze the dynamics behind attention sink that relates to MLLM hallucinations. An in-depth investigation into the instruction-tuning stage of training reveals that, attention sink appears at the turning point of image-text relevance in model responses (Section 3.1), which contains misleading global information that triggers subsequent hallucination content (Section 3.2).

3.1 Analysis of Instruction-tuning Datasets

To enable user conversation with both text and image inputs, MLLMs are first pre-trained on large-scale datasets containing image-text pairs [11], and then fine-tuned on instruction datasets tailored for downstream applications. The instruction-tuning datasets consist of task descriptions and task-specific input-output pairs, covering multi-modal tasks such as image captioning [7], visual question answering (VQA) [39, 46], and referring expression comprehension (REC) [8]. For example, LLaVA [46] uses GPT-4 [1] to generate instruction-following dialogues, by providing it with captions and bounding boxes of COCO [41] images. The resulting dataset, LLaVA-Instruct-150k, has been utilized to fine-tune MLLMs like LLaVA [46], Shikra [10], and InstructBLIP [16].

Although instruction-tuning datasets include fine-grained question-answering pairs, the text-image relevance in model responses shows a decreasing trend. A closer examination of the ground truth responses reveals that, after describing the

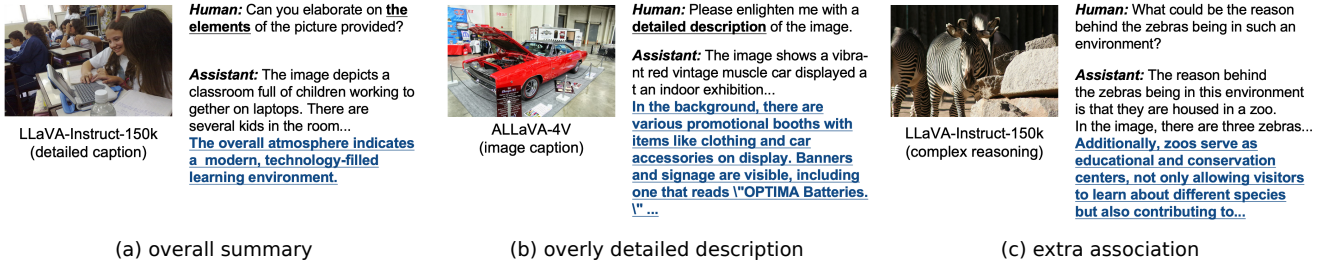


Figure 2: Examples of inferred elements in ground truth responses: (a) overall summary of the image content, (b) overly detailed description of trivial objects, and (c) extra association not instructed by the task. The texts after "Human:" denote instructions, and those after "Assistant:" are ground truth responses. The examples are selected from the LLaVA-Instruct-150k [46] and ALLaVA-4V [7] datasets of detailed image caption and complex reasoning tasks, which are generated with GPT-4 [1] and GPT-4V [55] models respectively.

image content and following the instructions, the responses generally include additional inferred elements, such as overall summaries, overly detailed descriptions, and extra associations based on the image content, as displayed in Fig. 2. This may be attributed to the fact that models like GPT-4 [1] and GPT-4V [55], which are used for data generation, have strong comprehension and associative abilities. As a result, they tend to offer extra references and details in a user-friendly manner.

To illustrate the decreasing text-image relevance in model responses of open-source instruction-tuning datasets, we select CLIPScore [26] as a metric. The CLIPScore is generally adopted to evaluate the image-text compatibility [65], which first extracts the embeddings for both visual and textual inputs with CLIP [60] model, and then calculates the cosine similarity between these embeddings to reveal their relevance. We compute the CLIPScore between the input images and each sentence in the ground truth responses of the LLaVA-Instruct-150k [46] and ALLaVA-4V [7] datasets. Fig. 3 reveals that, the ground truth responses exhibit a significant decrease in image-text relevance after the first few sentences. It results in two distinct segments in model-generated responses: (1) first the detailed descriptions closely tied to the image, and (2) content that is either loosely related to the image or beyond the visual interpretability of MLLMs.

The innate problem of datasets contributes to the hallucination problems of released MLLMs. When fine-tuned on such datasets, MLLMs tend to adopt the pattern of two-segment responses, first describing the image and then generating associative content. Moreover, when trained to fit the second part of the responses, MLLMs are compelled to generate details that they cannot visually comprehend [81], or abstract statements unrelated to the instructions. We also observe that the attention sink phenomenon emerges at the turning point of image-text relevance, which generally leads the hallucination responses with loose relation with images. We discover the following properties of attention sink originating from the instruction-tuning training:

(1) MLLMs inherit the *two-segment response* pattern from

instruction-tuning datasets. We prompt MLLMs to generate detailed image captions for VG 100K [31] dataset, and evaluate the per-sentence CLIPScore between input images and their responses, as shown in Fig. 4 (a)-(b). Similar to the trend observed in instruction-tuning datasets, the MLLM responses clearly show a significant decline in image-text relevance, which applies to all three decoding strategies.

(2) **Attention sink appears at the turning point of CLIP-Score.** By identifying the columnar patterns within the attention maps, we trace the presence of sink tokens and evaluate the mean CLIPScore of model responses before and after them, as shown in Fig. 4 (c). Our findings reveal that the attention sink appears to segment the response, with a marked decrease in image-text CLIPScore following the sink token, which suggests less relevant content and the prone to hallucinations. Notably, this issue is observed not only on models that are instruct-tuned on datasets displaying these tendencies (e.g., InstructBLIP and LLaVA-1.5 trained on LLaVA-Instruct-150k), but is also prevalent on MLLMs like MiniGPT-4, which are trained on closed-source datasets. This observation highlights a widespread problem across existing instruction-tuning paradigms.

3.2 Aggregated information in Attention Sink

To explain the emergence of attention sink at the turning point of image-text relevance, we dig deeper into the attention mechanism during MLLM generation. We notice that, besides the high attention scores and columnar patterns, sink tokens are predominantly non-content tokens (e.g., punctuation marks and article words) that convey minimal semantic meaning. For instance, in the responses of LLaVA-1.5, up to 73.5% of the sink tokens are non-content, indicating a tendency of allocating high attention to these semantically trivial elements.

We related this observation with a unique behavior discovered in Transformer-based models: *the aggregation of knowledge*. The process occurs when global information of inputs is aggregated into uninformative tokens, providing a

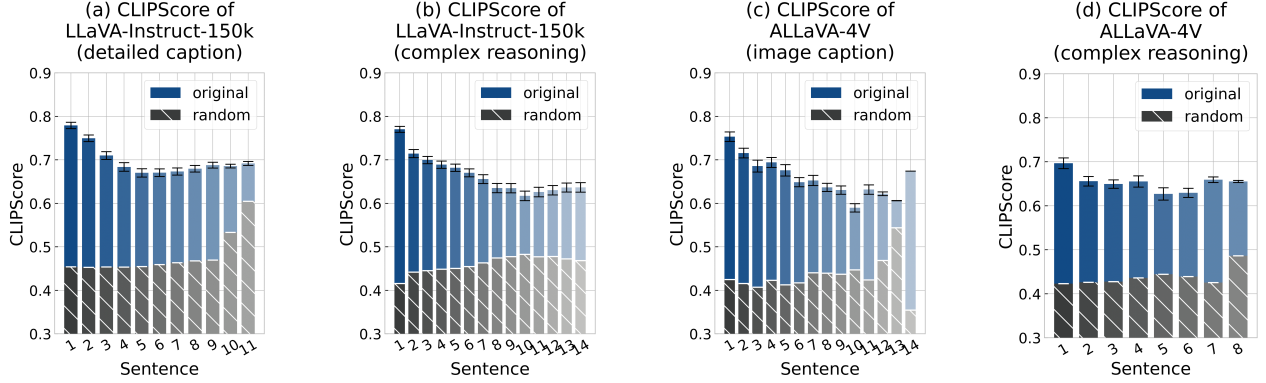


Figure 3: Per-sentence CLIPScore between input images and ground truth responses in instruction-tuning datasets. We report CLIPScore between input images and random response sentences as the baseline, denoted as *random*.

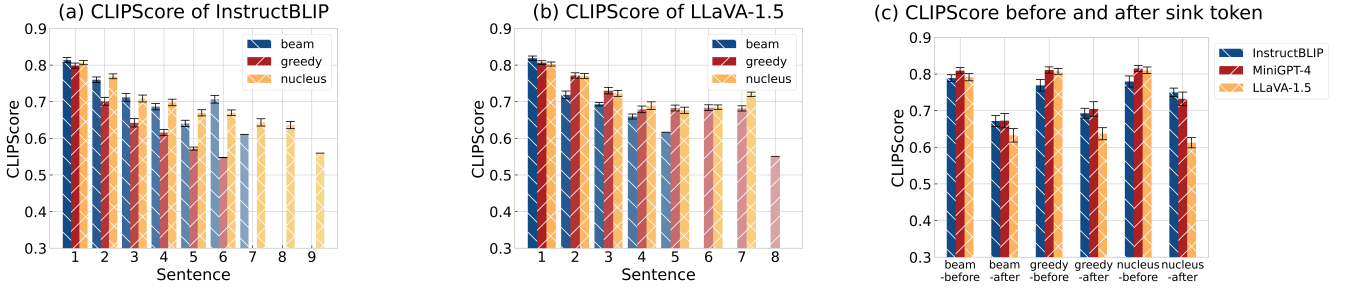


Figure 4: **(a)-(b)** Per-sentence CLIPScore between input images and MLLM responses of InstructBLIP and LLaVA-1.5. **(c)** Mean CLIPScore of MLLM responses before and after the sink token. The postfix *-beam*, *-greedy*, and *-nucleus* represent beam search, greedy search, and nucleus sampling decoding respectively. The missing bars indicate no generated sentences of the corresponding length.

shortcut for the subsequent generation or classification. The phenomenon is observed in Transformer-based models like Vision Transformers (ViTs) [18], LLMs, and MLLMs. For example, in language models, information is aggregated into functional label words (e.g., words like *positive* and *negative* in the task of sentiment analysis) in shallow layers to support final predictions [69]. Similarly, in ViTs, where image patches are treated as tokens, the models inject global information into some background tokens to replace their local information, which facilitates the training of linear models for classification [17]. In the study of MLLM hallucination, [28] also hypothesizes that certain tokens in MLLM responses aggregate crucial knowledge from contexts, and over-reliance on these tokens can lead to a neglect of the entire image content.

Leading by the common phenomenon of aggregating behaviors, we note that part of the global information in MLLM, representing visual and textual inputs, is also aggregated into sink tokens. Fig. 5 presents a distribution of cosine similarity between the middle-layer embeddings of multi-modal inputs and the generated tokens. It’s notable that sink tokens, which appear at the turning points of CLIPScore, exhibit a significantly higher resemblance to global input information compared to other tokens. We relate this observation to the

hallucinated generation, and make the following analysis.

(1) Attention sinks aggregate information as global context. The aggregating behavior of Transformer-based models is formed naturally during training, with sink tokens receiving high attention scores to aid in subsequent prediction or generation. In Fig. 5, the higher similarity to input embeddings indicates that global multi-modal information is partly integrated into the sink tokens. In the generation process of MLLMs, multi-modal input tokens are positioned before the entire response, serving as a global context. We hypothesize that, inheriting the two-segment response pattern (Section 3.1), attention sinks are chosen to distinguish between segments with different focus, content, and style in MLLM generation. This mechanism provides a more relevant global context for the latter part of the model’s responses, minimizing the need for long-distance attention and aligning with the observed MLLM generation patterns.

(2) Misleading aggregation triggers hallucinated response. While the aggregation process aligns with the generation pattern of MLLMs, we note that only part of the global information is fused into sink tokens, which deviates from the original global information. We speculate that it is still due to deficiencies in the instruction tuning phase, where the second part of

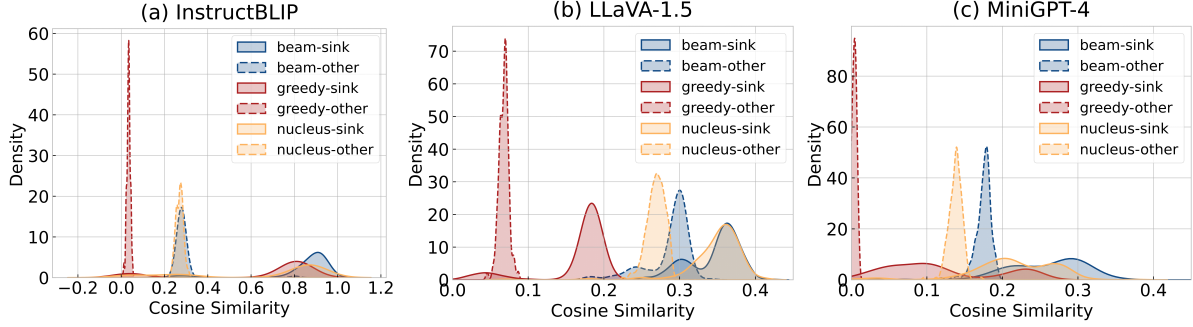


Figure 5: Distribution of cosine similarity between multi-modal input embeddings and generated token embeddings. We compare the similarity of sink tokens (with the postfix *-sink*) and all other tokens (with the postfix *-other*).

responses in training data often includes irrelevant descriptions (Section 3.1), and will mislead the aggregating process with partial, trivial, and even wrong global information. Furthermore, the aggregation of global context into a single token inevitably results in a significant loss of information, diminishing the factual accuracy of the image content. Consequently, MLLMs are trained to aggregate misleading information as context for irrelevant generations. The high attention scores assigned to these sink tokens exacerbate the hallucination problem, introducing irrelevant objects, confused attributes, and incorrect relationships.

4 Our Hallucination Attack

4.1 Motivation

Based on the overall analysis in Section 3, we highlight two important properties of attention sink in MLLM hallucination. First, the emergence of attention sink does not depend on specific textual or visual inputs, as the reason behind hallucination is deeply rooted in the attention mechanism during generation. Second, it requires no prior knowledge or external reference for detection, only the attention maps during the generation process.

To promote the safe, reliable, and beneficial development of MLLMs, we propose the first hallucination attack utilizing only the attention sink phenomenon, to explore the current state of MLLM hallucination severity and existing mitigation strategies. By manipulating the attention mechanism and hidden states during generation, our method achieves dynamic, effective, and highly transferrable attacks, all without the need for additional human intervention. The attacking strategy requires no pre-defined target responses, and applies to general visual and textual inputs, overcoming a major limitation of current adversarial attacks against MLLMs.

4.2 Security Settings

Attack Scenario. We define the goal of our hallucination attack against MLLMs as increasing the amount of hallucinated

content in MLLM responses, which includes inconsistencies in objects, attributes, and relationships according to the actual image content. The target models include open-sourced MLLMs and commercial MLLM APIs (such as GPT-4o provided by OpenAI). The target MLLMs may have built-in filtering mechanisms for multi-modal inputs or may be enhanced with additional mitigation strategies. The ultimate malicious goal can manifest in various forms, which include intensifying hallucination in subsequent conversations, delivering misleading information to users, and causing incorrect decisions in downstream applications, particularly in critical fields where even minor errors can have severe consequences. In real-world scenarios, many MLLM applications (e.g., AI assistants, customer service, physical therapy, and document analysis [75]) allow user-defined uploads, which are vulnerable to adversarial inputs. Additionally, MLLMs in black-box systems (e.g., autonomous driving) are also susceptible to threats like physical adversarial patches [27].

Threat Model. We conceive an attacker who exploits adversarial visual inputs to perform hallucination attacks. With white-box access to a surrogate MLLM, the attacker can compute gradients and construct adversarial examples. The attacker’s objective is to compromise the faithfulness of target MLLMs, thereby maliciously influencing their downstream applications. Driven by this goal, the attacker is highly motivated to deploy these adversarial inputs against black-box MLLMs and proprietary commercial MLLM APIs. To evade detection and filtering mechanisms employed by commercial platforms, the attacker must ensure that the adversarial inputs still produce high-quality and useful MLLM responses.

4.3 Method

4.3.1 Formulation of MLLM Generation

When processing multi-modal instructions during user interaction, the MLLMs take both visual and textual prompts as context. We denote the visual input tokens as $\mathbf{x}^v = \{x_0, x_1, \dots, x_{N-1}\}$, where N represents the length of visual inputs, and is typically predefined during training. The textual

input tokens are denoted as $\mathbf{x}' = \{x_N, x_{N+1}, \dots, x_{M+N-1}\}$, with an input length of M . The multi-modal inputs are concatenated into a single sequence $\mathbf{x}^{in} = \{x_i\}_{i=0}^{M+N-1}$, which is then fed into the MLLMs for further alignment and generation.

The response of MLLMs is sampled in an auto-regressive manner, with each token predicted based on previously generated ones. During the inference, the hidden states of each token are extracted as embeddings. We represent the hidden states of token i in the l -th layer as $h_i^{(l)}$, and the self-attention scores of each token as $a_i^{(l)}$, which is formulated as:

$$\mathbf{h}^{(l)} = \{h_0^{(l)}, h_1^{(l)}, \dots, h_{T-1}^{(l)}\} \in \mathbb{R}^{T \times d} \quad (1)$$

$$a_i^{(l)} = \text{Softmax} \left(\frac{Q^{(l)} K^{(l)T}}{\sqrt{d_k}} \right)_i \quad (2)$$

$$\mathbf{a}^{(l)} = \{a_0^{(l)}, a_1^{(l)}, \dots, a_{T-1}^{(l)}\} \in \mathbb{R}^{T \times T} \quad (3)$$

where T denotes the overall length of tokens including model responses, d denotes the dimension of hidden states, $Q^{(l)} = \mathbf{h}^{(l)} W_Q^{(l)} \in \mathbb{R}^{T \times d_k}$ and $K^{(l)} = \mathbf{h}^{(l)} W_K^{(l)} \in \mathbb{R}^{T \times d_k}$ represents query and key vectors with a dimension of d_k , after the linear projection of hidden states. In the last layer of MLLMs, a vocabulary head \mathcal{H} projects the final hidden states $\mathbf{h}^{(L)}$ into probabilities, which supports the next-token prediction.

$$x_{<t} = \{x_0, x_1, \dots, x_{t-1}\} \quad (4)$$

$$p(x_t | x_{<t}) = \text{Softmax}(\mathcal{H}(\mathbf{h}^{(L)}))_{x_t} \quad (5)$$

where $M+N \leq t < T$, and $x_t \in \mathcal{X}$ in which \mathcal{X} means the whole vocabulary set. With the probability of the next token $p(x_t | x_{<t})$, different decoding strategies such as beam search, greedy search, and nucleus sampling are adopted to output the model responses. After predicting the t -th token, it will be appended at the end of the token list for the next-round generation. This process continues until the model predicts an end-of-sentence (EOS) token, signaling the end of responses.

4.3.2 Identifying Potential Sink Tokens

According to our analysis in Section 3, the attention sink appears at the turning point of image-text relevance, which aggregates some misleading global information from multi-modal inputs, and provokes hallucinated content.

When conducting the hallucination attack with adversarial visual inputs, we aim to coax the target MLLMs into generating as many attention sinks during their responses, which has been demonstrated to significantly increase the hallucinated content. Given that the maximum length of model responses and the decoding strategies cannot be fully covered in the attacking process, the adversarial perturbations are demanded to include the generalized patterns that trigger more sink tokens in MLLMs generation, rather than merely producing sink tokens at fixed token positions.

To identify the potential tokens that are most likely to become attention sinks, we perform a search in current model responses, based on the amount of global information contained in each token. We retrieve the hidden states of tokens $\mathbf{h}^{(l)}$ in the intermediate layer of MLLMs, and compute the cosine similarity between the global input tokens \mathbf{x}^{in} and each generated token x_i as follows:

$$s_i^{(l)} = \text{Sim}(h_i^{(l)}, \bar{h}_g^{(l)}), 1 \leq l < L, i > M+N-1 \quad (6)$$

where $\text{Sim}(\cdot, \cdot)$ represents the computation of cosine similarity, $\bar{h}_g^{(l)}$ is defined as the mean vector of the l -th layer hidden states of multi-modal inputs, i.e., $\bar{h}_g^{(l)} = \text{Mean}(\{h_i^{(l)}\}_{i=0}^{M+N-1})$.

Based on the global information each token absorbs within the middle-layer embeddings, we choose the token index with the highest $s_i^{(l)}$ as the potential sink position in the subsequent optimization of adversarial perturbation:

$$\text{idx} = I[\max\{s_{M+N}^{(l)}, s_{M+N+1}^{(l)}, \dots, s_{T-1}^{(l)}\}] \quad (7)$$

where $I[\cdot]$ indicates the index of the token with the maximum similarity score.

In each round of adversarial optimization, we will identify the potential token according to current model responses. As the sentence structure remains stable in consecutive attack rounds, the selection of potential tokens tends to be focused over a few iterations, leading to concentrated and targeted optimization towards sink tokens.

4.3.3 Optimizing Adversarial Perturbations

After searching for the potential token to induce attention sink, we design the adversarial target based on the ideal characteristics of sink tokens, which covers both attention behavior and knowledge aggregation. Introducing sink tokens with high attention scores, we demand that the subsequent generation includes declined image-text relevance, and produce hallucinated content from misleading information.

Attention Loss. When constructing the desired columnar attention behavior, we acquire the middle-layer attention scores within MLLM generation, which is $\mathbf{a}^{(l)} = \{a_i^{(l)}\}_{i=0}^{T-1}$. The attention score of each token $a_i^{(l)} \in \mathbb{R}^i, 0 \leq i < T$ contains its allocated attention on the previous sequence. We construct the attention map $A^{(l)} = [a_0^{(l)}, a_1^{(l)}, \dots, a_{T-1}^{(l)}] \in \mathbb{R}^{T \times T}$ with obtained scores, where $a_{i,j}^{(l)}$ represents the attention scores of token x_i allocated to x_j .

In the attention map $A^{(l)}$, a columnar attention pattern on certain tokens x_i means that the subsequent tokens $x_j, j > i$ all allocate high attention to the sink token, and make predictions with its domination. To induce such a columnar attention pattern on the chosen potential token x_{idx} , we focus on a localized attention window $A^{(l)} = [a_{\text{idx}}^{(l)}, a_{\text{idx}+1}^{(l)}, \dots, a_{T-1}^{(l)}]$, and

the attention loss is defined as:

$$\mathcal{L}_{\text{attn}}(\mathbf{x}^v, \mathbf{x}^t) = CE(A^{(l)}, \text{idx}) \quad (8)$$

where $CE(\cdot)$ denotes the cross entropy loss function. The attention loss forces subsequent tokens to allocate high attention to the potential sink token, thereby forming the attention sink adversarially to introduce a decline of image-text relevance.

Embedding Loss. Besides the direct phenomenon of columnar attention behaviors, our observation also suggests that sink tokens bear a much higher resemblance to the multi-modal inputs, which aggregate misleading global information originating from the instruction tuning. In the hallucination attack, we aim to increase the cosine similarity of the embeddings between potential sink tokens and global input information. The hidden states of potential token in the l -th intermediate layer $h_{\text{idx}}^{(l)}$ is obtained, and is used to compute its similarity of global information $\text{Sim}(h_{\text{idx}}^{(l)}, \bar{h}_g^{(l)})$.

Since the aggregation of global information encourages the formation of an attention sink, we also introduce an embedding loss to partially raise the global information embedded in the potential tokens. We adopt the hinge loss to ensure the embedding only contains incomplete and misleading information:

$$\mathcal{L}_{\text{emb}}(\mathbf{x}^v, \mathbf{x}^t) = \max(0, \sigma - \text{Sim}(h_{\text{idx}}^{(l)}, \bar{h}_g^{(l)})) \quad (9)$$

where σ is a hyper-parameter predefined with the observation of sink tokens in different MLLMs. The embedding loss is designed to promote the injection of more misleading global information into the potential token, which helps MLLMs spontaneously generate hallucinatory content in subsequent outputs.

Adversarial Objective. With the hallucination attack targeting both attention behavior and hidden states, the overall adversarial objective is defined as:

$$\begin{aligned} \min \quad & \mathcal{L}_{\text{attn}}(\tilde{\mathbf{x}}^v, \mathbf{x}^t) + \alpha \mathcal{L}_{\text{emb}}(\tilde{\mathbf{x}}^v, \mathbf{x}^t) \\ \text{s.t.,} \quad & \tilde{\mathbf{x}}^v = \mathbf{x}^v + \delta, \|\delta\|_p < \varepsilon \end{aligned} \quad (10)$$

where δ is the adversarial perturbation on the visual input \mathbf{x}^v , α is the hyper-parameter to adjust the regularization, $\|\cdot\|_p$ is the computation of p-norm, and ε is the attack budget which controls the magnitude of the adversarial perturbation.

The hallucination attack will be conducted in multiple rounds, where the adversarial objective will direct the optimization of adversarial perturbation in each round. The attacking algorithm of our method is illustrated in Alg. 1.

5 Experiments

5.1 Experimental Settings

Target Models. To conduct a comprehensive evaluation of mainstream MLLMs in the open-source community, we select

Algorithm 1 Hallucination Attack

Input: The visual input: \mathbf{x}^v , the textual input: \mathbf{x}^t , and the target model: $mllm$
Output: The adversarial visual input: $\tilde{\mathbf{x}}^v$

- 1: $S \leftarrow 0$.
- 2: **while** $S < S_{\text{max}}$ **do**
- 3: $response \leftarrow mllm(\tilde{\mathbf{x}}^v, \mathbf{x}^t)$.
- 4: Get hidden states $\mathbf{h}^{(l)}$ from the l -th layer.
- 5: Get attention scores $\mathbf{a}^{(l)}$ from the l -th layer.
- 6: $\bar{h}_g^{(l)} \leftarrow \text{Mean}(h_0^{(l)}, h_1^{(l)}, \dots, h_{M+N-1}^{(l)})$.
- 7: Compute the cosine similarity between $\bar{h}_g^{(l)}$ and generated token embeddings as $s_i^{(l)} = \text{Sim}(h_i^{(l)}, \bar{h}_g^{(l)})$
- 8: Get potential token idx = $I[\max\{s_{M+N}^{(l)}, \dots, s_{T-1}^{(l)}\}]$.
- 9: $A^{(l)} \leftarrow$ construct attention map with $\mathbf{a}^{(l)}$.
- 10: $\mathcal{L}_{\text{attn}} \leftarrow CE(A^{(l)}, \text{idx})$. ▷ See Equation 8
- 11: $\mathcal{L}_{\text{emb}} \leftarrow \max(0, \sigma - \text{Sim}(h_{\text{idx}}^{(l)}, \bar{h}_g^{(l)}))$. ▷ See Equation 9
- 12: Compute gradient $g \leftarrow \nabla_{\tilde{\mathbf{x}}^v} \mathcal{L}_{\text{attn}} + \alpha \mathcal{L}_{\text{emb}}$
- 13: Updating $\tilde{\mathbf{x}}_S^v \leftarrow \tilde{\mathbf{x}}_S^v - \gamma \cdot \text{sign}(g)$.
- 14: Clipping $\tilde{\mathbf{x}}_S^v \leftarrow \text{Clip}(\tilde{\mathbf{x}}_S^v, -\varepsilon, \varepsilon)$.
- 15: $S \leftarrow S + 1$
- 16: **end while**
- 17: **return** $\tilde{\mathbf{x}}^v$

four of the most representative MLLMs including Instruct-BLIP [16], MiniGPT-4 [86], LLaVA-1.5 [45], and Shikra [10] as target models. The details about visual encoders and LLM backbones are available in Tab. 1, which includes well-trained vision models like EVA [19] and CLIP [60], and widely-used LLMs like Vicuna [14] and LLaMA [66]. To evaluate the adversarial effects on closed-source commercial APIs, we also take the recently released GPT-4o mini [57] and Gemini 1.5 flash [56] into consideration. More implementation details are available in Appendix A.

MLLM Tasks. To comprehensively evaluate the adversarial impact of the hallucination attack, we focus on two types of downstream tasks for MLLMs: image captioning and question-answering (QA). These tasks represent different aspects of MLLMs' open-ended generation capabilities.

Metrics. We assume that the attacker's goal is to intensify hallucinations in MLLM responses without sacrificing their quality and helpfulness. For the image captioning task, we evaluate both the extent of hallucination and the quality of the generated responses. For the QA task, we evaluate the accuracy of model answers.

Evaluation of hallucination. In assessing the severity of object hallucination, earlier research adopted the Caption Hallucination Assessment with Image Relevance (CHAIR) metric [61]. However, this metric fails to consider the hallucination of attributes and relationships, and only supports closed-ended evaluation (i.e., covering only 80 object classes in MS-COCO [41] datasets). To achieve more advanced evaluation

Table 1: Basic information of open-source MLLMs in our experiments.

MLLM	InstructBLIP	MiniGPT-4	LLaVA-1.5	Shikra
Visual Encoder	Vicuna-7b-v1.1	Vicuna-7b-v0	Vicuna-7b	LLaMA-7b
LLM Backbone	EVA-ViT-g/14	EVA-ViT-g/14	CLIP-ViT-L/14	CLIP-ViT-L/14

for open-ended model responses, we follow the previous studies [28, 44] and conduct GPT-4 assisted evaluation based on HalluBench [83]. The benchmark is based on a subset of the Visual Genome (VG) dataset [31], featuring images annotated with bounding box coordinates and detailed descriptions of objects, attributes, and relationships. The HalluBench also provides human-annotated factual information to improve evaluation accuracy. During the evaluation, the GPT-4 model is prompted with all the references and judges the hallucination in MLLM responses sentence by sentence.

Evaluation of response quality. The overall quality of the MLLM’s responses is critical in real-world attacking scenarios. If the adversarial visual inputs result in low-quality responses characterized by poor fluency, distorted sentence structure, or excessive grammatical errors, downstream users of MLLMs can easily detect these patterns and implement filtering mechanisms. In the evaluation, we also resort to the GPT-4 model to assess the grammar, fluency, and naturalness of generated responses. The model provides a score between 0-9 as the overall response quality.

Accuracy of QA. To evaluate the attacking effects on the QA task, we use the OK-VQA [52] benchmark, which is based on the MS-COCO image dataset and widely recognized for assessing the general QA capabilities of MLLMs [3, 10]. We select a subset of the OK-VQA benchmark, consisting of questions paired with 5 human-annotated answers each. For evaluation, we employ the standard VQA accuracy metric [2], which measures model responses by comparing them to ground truth answers. A response is considered 100% accurate if at least three human annotators provide the exact same answer. The VQA accuracy metric is defined as:

$$\text{accuracy} = \min\left(\frac{\# \text{ humans that provided that answer}}{3}, 1\right) \quad (11)$$

5.2 Adversarial Effects

Image Captioning Task. To investigate the impact of adversarial visual inputs, we first concentrate on the level of hallucination in white-box models. During the hallucination attack, we construct adversarial visual inputs with the attack budget ϵ of 2/255, 5/255, and 8/255, to observe the adversarial effects under different strengths of perturbations. During evaluation, we follow the setting of HalluBench, and query all MLLMs with a text prompt of *Please describe this image in detail.*, together with the constructed adversarial visual inputs.

The adversarial effects on white-box target MLLMs with beam search decoding are reported in Tab. 2. Remarkably, our

hallucination attack induces a substantial amount of hallucinated content in MLLM responses, achieving up to 75.74% hallucinated words in single responses with only minor perturbations on visual inputs. With larger attack budgets, the number of affected sentences and words has noticeably increased. In the real-world applications of MLLMs, the severer hallucination in model responses may mislead downstream modules to focus on inaccurate descriptions of visual content and crash on making planning or decisions. It is also worth noting that, our hallucination attack does not work by increasing the output length (with similar SPI and WPI in results). This contrasts with the approach in [22] that delaying the occurrence of (EOS) token may lead to an increase in CHAIR metrics. However, their eight times longer model responses are easy to be detected from normal uses, and inevitably affect their fluency and helpfulness.

During the attack process, we observe that as the adversarial noise was continuously optimized, changes in the sink token and MLLM responses align with our analysis in Section 3. Fig. 6 visualizes the attack process and the affected model response. During the optimization, the sink token changes from tokens with concrete meaning (e.g., *item* in step 1) to non-content ones (e.g., the comma , in step 18). The manipulation of attention behaviors also leads to more hallucinated objects (e.g., *cup* and *woman*) and wrong relationships (e.g., *holding a bowl in his hand*) during the attack process. With adversarial visual inputs, the target model tends to make up contents not aligned with images, with an obvious attention sink phenomenon observed in the attention map.

QA Task. The white-box adversarial effects on the OK-VQA benchmark is presented in Tab. 4. The results show a reduction in accuracy of up to 7.67%, emphasizing that adversarial visual inputs significantly impair the general visual understanding capabilities of MLLMs. As a result, hallucinated answers may be generated and presented to downstream users, potentially leading to the propagation of misconceptions in real-world applications.

To better understand the adversarial effect on model responses, we also engaged human experts to classify the hallucinations induced by our attack, as presented in Fig. 9.

5.3 Black-box Transferability

In real-world attack scenarios, the target MLLMs usually remain inaccessible to the attackers. To demonstrate the effectiveness of hallucination attack under such settings, we

Table 2: Results of GPT-4 assisted hallucination evaluation for the image captioning task on white-box models. All of the MLLM responses are generated with *beam search* decoding. We report six aspects of evaluation, including the number of sentences per image (**SPI**), the number of words per image (**WPI**), the number of hallucinated sentences per image (**HSPI**), the number of hallucinated words per image (**HWPI**), the ratio of hallucinated sentences (**HSR**), and the ratio of hallucinated words (**HWR**). A larger HSPI, HWPI, HSR, and HWR indicate a higher level of hallucination in MLLM responses. The best results are marked in bold, and the number in brackets indicates the hallucination improvement compared to the clean image.

Target Model	Input	SPI	WPI	HSPI	HWPI	HSR(%)	HWR(%)
InstructBLIP	clean image	4.54	75.64	2.83	48.05	62.91%	64.93%
	$\epsilon=2/255$	4.60	80.19	2.97 (+0.14)	55.14 (+7.09)	64.92% (+2.01%)	68.23% (+3.30%)
	$\epsilon=5/255$	4.47	80.48	3.04 (+0.21)	54.90 (+6.85)	68.41% (+5.50%)	70.84% (+5.91%)
	$\epsilon=8/255$	4.41	79.71	2.89 (+0.06)	52.91 (+4.86)	66.79% (+3.88%)	69.45% (+4.52%)
LLaVA-1.5	clean image	4.60	116.24	2.68	79.08	59.62%	71.68%
	$\epsilon=2/255$	4.64	96.60	2.76 (+0.08)	62.97 (-16.11)	60.26% (+0.64%)	68.17% (-3.51%)
	$\epsilon=5/255$	4.49	108.03	2.67 (-0.01)	74.85 (-4.23)	62.36% (+2.74%)	75.74% (+4.06%)
	$\epsilon=8/255$	4.53	103.58	2.92 (+0.24)	75.45 (-3.63)	65.07% (+5.45%)	75.08% (+3.40%)
MiniGPT-4	clean image	3.98	60.56	2.31	37.34	58.13%	62.77%
	$\epsilon=2/255$	4.10	59.20	2.49 (+0.18)	37.97 (+0.63)	61.42% (+3.29%)	65.01% (+2.24%)
	$\epsilon=5/255$	3.97	66.27	2.41 (+0.10)	43.48 (+6.14)	61.02% (+2.89%)	67.09% (+4.32%)
	$\epsilon=8/255$	4.00	64.51	2.55 (+0.24)	40.83 (+3.49)	64.59% (+6.46%)	67.97% (+5.20%)
Shikra	clean image	3.11	46.13	1.56	23.39	52.95%	53.16%
	$\epsilon=2/255$	3.13	45.99	1.69 (+0.13)	25.65 (+2.26)	56.04% (+3.09%)	57.93% (+4.77%)
	$\epsilon=5/255$	3.26	46.82	1.83 (+0.27)	26.51 (+3.12)	57.88% (+4.93%)	58.25% (+5.09%)
	$\epsilon=8/255$	3.12	45.19	1.69 (+0.13)	25.69 (+2.30)	56.31% (+3.36%)	59.11% (+5.95%)

Table 3: Results of GPT-4 assisted hallucination evaluation for the image captioning task on black-box models. All of the MLLM responses are generated with *beam search* decoding. The six aspects of evaluation are the same as in Tab. 2. A larger HSPI, HWPI, HSR, and HWR indicate a higher level of hallucination in MLLM responses. The best results are marked in bold, and the number in brackets indicates the hallucination improvement compared to the clean image for each target model.

Surrogate Model	Target Model	SPI	WPI	HSPI	HWPI	HSR(%)	HWR(%)
InstructBLIP	InstructBLIP	4.47	80.48	3.04 (+0.21)	54.90 (+6.85)	68.41% (+5.50%)	70.84% (+5.91%)
	LLaVA-1.5	4.46	99.77	2.64 (-0.04)	70.51 (-8.57)	59.42% (-0.20%)	71.48% (-0.20%)
	MiniGPT-4	3.84	63.00	2.31	40.54 (+3.20)	61.81% (+3.68%)	68.21% (+5.44%)
	Shikra	3.20	48.95	1.79 (+0.23)	27.77 (+4.38)	56.14% (+3.19%)	57.09% (+3.93%)
LLaVA-1.5	LLaVA-1.5	4.49	108.03	2.67 (-0.01)	74.85 (-4.23)	62.36% (+2.74%)	75.74% (+4.06%)
	InstructBLIP	4.47	78.31	2.81 (-0.02)	51.85 (+3.80)	65.37% (+2.46%)	68.75% (+3.82%)
	MiniGPT-4	3.95	63.60	2.32 (+0.01)	42.25 (+4.91)	60.79% (+2.66%)	68.14% (+5.37%)
	Shikra	3.08	45.94	1.94 (+0.38)	29.96 (+6.57)	63.85% (+10.90%)	65.90% (+12.74%)
MiniGPT-4	MiniGPT-4	4.00	64.51	2.55 (+0.24)	40.83 (+3.49)	64.59% (+6.46%)	67.97% (+5.20%)
	InstructBLIP	4.36	79.62	2.96 (+0.13)	54.82 (+6.77)	68.96% (+6.05%)	71.94% (+7.01%)
	LLaVA-1.5	4.27	116.50	2.51 (-0.17)	75.86 (-3.22)	60.84% (+1.22%)	73.67% (+1.99%)
	Shikra	3.33	49.67	1.92 (+0.36)	28.99 (+5.60)	58.86% (+5.91%)	59.61% (+6.45%)
Shikra	Shikra	3.12	45.19	1.69 (+0.13)	25.69 (+2.30)	56.31% (+3.36%)	59.11% (+5.95%)
	InstructBLIP	4.48	80.36	2.90 (+0.07)	54.39 (+6.34)	67.48% (+4.57%)	70.28% (+5.35%)
	LLaVA-1.5	4.43	110.61	2.71 (+0.03)	75.34 (-3.74)	64.77% (+5.15%)	77.35% (+5.67%)
	MiniGPT-4	3.97	72.35	2.35 (+0.04)	46.90 (+9.56)	60.86% (+2.73%)	70.07% (+7.30%)

Table 4: Results of QA accuracy on the OK-VQA benchmark. All of the MLLM answers are generated with *greedy* decoding. A lower accuracy indicates a higher level of hallucination in MLLM answers. The best results are marked in bold, and the number in brackets indicates the hallucination improvement compared to the clean image.

Target Model	Input	Surrogate Model			
		InstructBLIP	LLaVA-1.5	MiniGPT-4	Shikra
InstructBLIP	clean image			56.33	
	$\epsilon=2/255$	51.33 (-5.00)	55.99 (-0.34)	52.33 (-4.00)	56.66 (+0.33)
	$\epsilon=5/255$	48.99 (-7.34)	53.99 (-2.34)	50.33 (-6.00)	50.33 (-6.00)
	$\epsilon=8/255$	49.66 (-6.67)	49.99 (-6.34)	47.33 (-9.00)	52.33 (-4.00)
LLaVA-1.5	clean image			60.00	
	$\epsilon=2/255$	56.66 (-3.34)	57.33 (-2.67)	57.66 (-2.34)	56.99 (-3.01)
	$\epsilon=5/255$	55.99 (-4.01)	56.66 (-3.34)	56.33 (-3.67)	52.33 (-7.67)
	$\epsilon=8/255$	54.66 (-5.34)	52.33 (-7.67)	54.99 (-5.01)	58.66 (-1.34)
MiniGPT-4	clean image			40.66	
	$\epsilon=2/255$	42.33 (+1.67)	38.33 (-2.33)	39.33 (-1.33)	42.33 (+1.67)
	$\epsilon=5/255$	39.33 (-1.33)	43.66 (+3.00)	36.66 (-4.00)	40.33 (-0.33)
	$\epsilon=8/255$	39.33 (-1.33)	43.33 (+2.67)	34.00 (-6.66)	39.99 (-0.67)
Shikra	clean image			55.33	
	$\epsilon=2/255$	55.00 (-0.33)	54.66 (-0.67)	56.99 (+1.66)	56.66 (+1.33)
	$\epsilon=5/255$	56.66 (+1.33)	54.00 (-1.33)	53.33 (-2.00)	51.33 (-4.00)
	$\epsilon=8/255$	51.66 (-3.67)	52.99 (-2.34)	52.66 (-2.67)	53.33 (-2.00)

construct adversarial visual inputs on one surrogate MLLM, and evaluate the transferability to both black-box MLLMs and closed-source commercial APIs. The attack is more challenging since the model structure, parameters and training paradigms of target MLLMs are unknown to the attackers. Commercial API providers such as OpenAI may also implement advanced defenses against multi-modal inputs [77], which have successfully blocked semantics-based visual attacks [48] and typographic attacks [24].

5.3.1 Attacking Black-box MLLMs.

Image Captioning Task. In our experimental setup, we select one of the four target MLLMs in Section 5.1 as the surrogate model, and evaluate the black-box transferability on the remaining ones. The results of hallucination evaluation for the image captioning task are presented in Tab. 3. Our proposed attack demonstrates high transferability across different structures of MLLMs, with the perturbed visual inputs achieving a 10.90% HSR and 12.74% HWR increase on black-box target models, even surpassing the increase in white-box attacks. We also find that the attack achieves better transferability on MLLMs with the same visual encoder architecture, possibly due to the similar distribution of hidden states.

QA Task. The black-box evaluation results for the QA task are reported in Tab. 4, demonstrating strong transferability across different architectures. The occasional increases in the accuracy of adversarial visual inputs may be attributed to the limitations of the manually labeled answers.

5.3.2 Attacking Closed-source Commercial APIs.

Image Captioning Task. As for attacking commercial APIs, our goal is to examine whether the proposed attack could evade potential defense and detection measures implemented by IT giants. We choose the latest commercial MLLM APIs, i.e., the GPT-4o mini [57] and Gemini 1.5 flash [56], as the target APIs. The adversarial transferability result for image captioning is displayed in Fig. 7. Though potential defenses are applied, the proposed attack still achieves a 3.40% and 5.32% increase in hallucination words. As the captions of the commercial MLLM APIs have been utilized in applications like medical diagnosis [71], science education [32], and financial decisions [40], the inaccurate interpretation induced by adversarial visual inputs deserves attention.

QA Task. In the QA task, the adversarial effects in Fig. 10 also reveal degraded visual comprehension and question-answering capabilities. Since commercial MLLM APIs are often regarded as powerful domain experts to guide the fine-tuning process (e.g., in medical [43, 64] and financial [20, 73] domains), our attack may significantly undermine the reliability of domain-specific MLLMs.

5.4 The Quality of Model Responses

When the attacker injects the adversarial visual input into target MLLMs, the model responses should resemble those in normal conversations, with no noticeable decline in quality. Otherwise, the stealth of the hallucination attack cannot

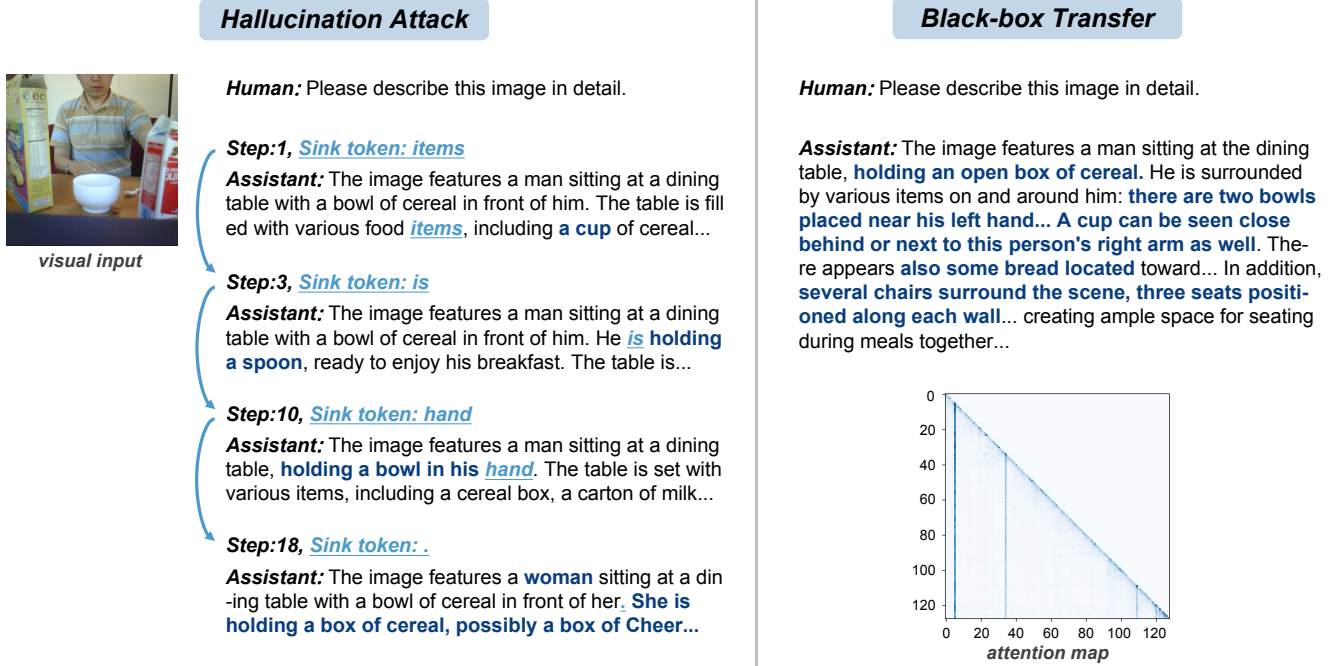


Figure 6: **Left:** The visualization of sink tokens and model responses during the hallucination attack. We display an optimization process of LLaVA-1.5 on the HalluBench dataset. **Right:** The black-box transfer effect of the constructed adversarial visual input on InstructBLIP. A pronounced attention sink phenomenon is observed in the attention map. The attention sink in model responses is denoted with **blue tokens**, and the hallucinated content is denoted with **indigo ones**.

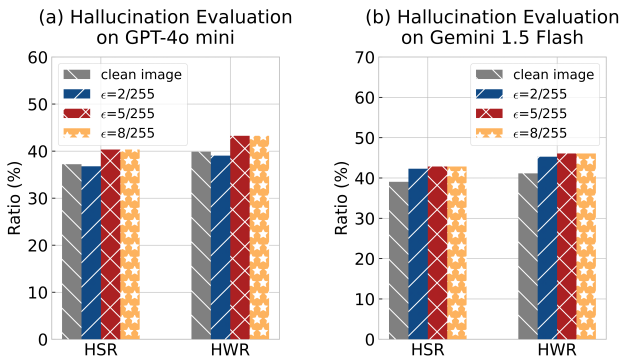


Figure 7: Results of GPT-4 assisted hallucination evaluation for image captioning task on commercial APIs: (a) GPT-4o mini released by OpenAI and (b) Gemini 1.5 Flash launched by Google. A larger HSR and HWR indicate more hallucinations in MLLM responses.

be ensured. We report the results of response quality for the image captioning task based on GPT-4 assisted evaluation in Fig. 8, and the results based on the Perplexity metric (PPL) in Fig. 11. Even if the model responses include more hallucinated content, they still maintain a high level of semantic accuracy, usefulness, and fluency, indicating the feasibility and stealthiness of our attack in practical settings.

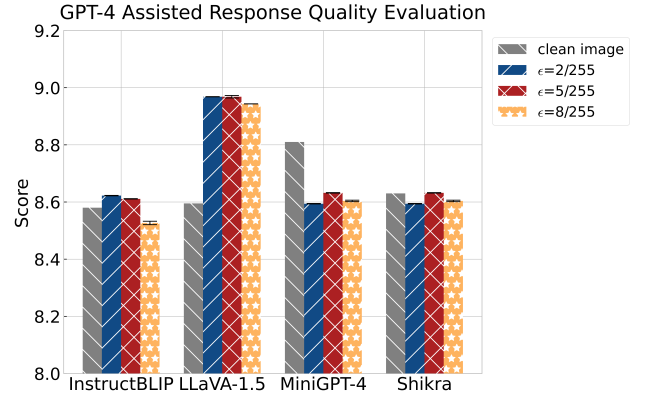


Figure 8: Results of GPT-4 assisted response quality evaluation, covering both white-box and black-box attack scenarios. A higher score reflects better MLLM response quality.

5.5 Attacking Mitigation Mechanisms

As discussed in Section 2.2, recent mitigation mechanisms of MLLM hallucination fall into three categories: mitigation with *decoding*, *model retraining*, and *post-processing*. To verify the effectiveness of our attack on existing defenses, we select the representative methods as follows. All experiments are performed with the official implementation and released

checkpoints in default parameters.

OPERA [28]. Targeting the attention sink phenomenon, this method detects the columnar behavior in MLLM’s self-attention matrices, and applies penalty terms to the model logits during the beam search inference. As a decoding-based mitigation approach, it directly blocks the attention behaviors associated with hallucination attacks, establishing a strong baseline for mitigation.

VCD [35]. Aiming to mitigate the inherent language prior, this method manipulates the decoding strategies of MLLMs, and adjusts the output logits with distorted visual inputs, which reflects the model’s reliance on statistical bias.

Less is more [81]. Noticing the overly detailed descriptions in MLLM instruction-tuning datasets, this method fine-tunes released MLLMs with additional supervision on EOS decisions, encouraging the models to stop generation timely before hallucinated content.

LRV-Instruction [44]. Addressing the limitation of existing instruction-tuning datasets that only contain samples of positive answers, this method constructs a comprehensive dataset with negative samples to guide models in identifying incorrect instructions. The dataset is used to fine-tune released MLLMs to mitigate hallucination.

LURE [85]. With a post-hoc rectification strategy, this method collects a hallucinatory dataset and fine-tunes another MLLM as the hallucination revisor. During inference, the trained revisor detects underlying object hallucinations with output logits and rectifies them with new responses.

Attacking Mitigation Mechanisms. We assess the adversarial effects on LLaVA-1.5 and MiniGPT-4 with the representative mitigation methods, and the results are detailed in Tab. 5 and Tab. 6 respectively. Experimental results demonstrate that the proposed attack succeeded in breaking all mitigation mechanisms in our evaluation, bringing the hallucination rate back to the level when no defenses are in place (e.g., the HWR achieves 75.54% in attacking LRV-Instruction defense, outperforming the one on vanilla model). This indicates the efficacy of our attack in bypassing existing mitigation methods, including adaptive strategies like OPERA. It also reveals the shortcomings of current mitigation strategies in defending deliberately crafted perturbations.

Attacking Adaptive Mitigation. Since the emergence of sink tokens triggers hallucinated content, we consider an early-stopping adaptive mitigation. With white-box access to the target MLLMs, this mitigation detects the attention sink phenomenon during generation and terminates output before sink tokens appear. Tab. 9 reports the length, quality, and hallucination metrics of MLLM responses under the mitigation strategy. Despite the implementation of adaptive mitigation, the adversarial visual inputs continue to provoke more severe hallucinations in model outputs. While adaptive mitigation reduces hallucinated content, it significantly decreases the mean length and quality of responses by 45.36% and 63.67% respectively, resulting in incomplete and less detailed descriptions

Table 5: Results of GPT-4 assisted hallucination evaluation against mitigation mechanisms on LLaVA-1.5. (*), (°), and (°) denote methods through *decoding*, *model retraining*, and *post-processing* respectively. Best results are marked in bold.

Mitigation	Input	HSR(%)	HWR(%)
OPERA * [28]	clean	50.27%	51.93%
	$\epsilon=2/255$	53.50% (+3.23%)	56.13% (+4.20%)
	$\epsilon=5/255$	52.33% (+2.06%)	54.37% (+2.44%)
	$\epsilon=8/255$	55.86% (+5.59%)	58.18% (+6.25%)
VCD * [35]	clean	51.38%	53.58%
	$\epsilon=2/255$	54.46% (+3.08%)	57.02% (+3.44%)
	$\epsilon=5/255$	57.69% (+6.31%)	60.12% (+6.54%)
	$\epsilon=8/255$	62.42% (+11.04%)	64.95% (+11.37%)
Less is More ° [81]	clean	43.74%	45.78%
	$\epsilon=2/255$	46.22% (+2.48%)	48.23% (+2.45%)
	$\epsilon=5/255$	47.68% (+3.94%)	49.91% (+4.13%)
	$\epsilon=8/255$	52.77% (+9.03%)	54.07% (+8.29%)

Table 6: Results of GPT-4 assisted hallucination evaluation against mitigation mechanisms on MiniGPT-4. (*), (°), and (°) denote methods through *decoding*, *model retraining*, and *post-processing* respectively. Best results are marked in bold.

Mitigation	Input	HSR(%)	HWR(%)
OPERA * [28]	clean	43.71%	45.79%
	$\epsilon=2/255$	57.09% (+13.38%)	59.34% (+13.55%)
	$\epsilon=5/255$	60.78% (+17.07%)	63.75% (+17.96%)
	$\epsilon=8/255$	59.03% (+15.32%)	61.82% (+16.03%)
LRV-Instruction ° [44]	clean	67.19%	70.82%
	$\epsilon=2/255$	69.73% (+2.54%)	73.81% (+2.99%)
	$\epsilon=5/255$	70.75% (+3.56%)	75.03% (+4.21%)
	$\epsilon=8/255$	71.43% (+4.24%)	75.54% (+4.72%)
LURE ° [85]	clean	48.57%	53.54%
	$\epsilon=2/255$	58.21% (+9.64%)	64.44% (+10.90%)
	$\epsilon=5/255$	59.42% (+10.85%)	67.41% (+13.87%)
	$\epsilon=8/255$	59.97% (+11.40%)	67.85% (+14.31%)

of image content. In real-world applications, this strategy may also incur substantial computational overhead and severely degrade the user experience for commercial APIs.

5.6 Baseline Comparison

To demonstrate the superiority of our attack in constructing adversarial perturbations that result in a severe level of hallucination, we consider visual inputs with random Gaussian noises as a baseline. The magnitude of random perturbation is set as the same of hallucination attacks. The attack effects of baseline methods with beam search decoding are displayed in Tab. 7. It is obvious that trivial random perturbation, though disrupting the feature extraction process of the visual encoder and enhancing the model’s reliance on language priors,

Table 7: Results of GPT-4 assisted hallucination evaluation of the baseline method with *beam search* decoding. δ denotes of budget of random noises injected into visual inputs. The line of *attack* denotes the best results in the white-box attack scenario.

	InstructBLIP		LLaVA-1.5		MiniGPT-4		Shikra	
	HSR(%)	HWR(%)	HSR(%)	HWR(%)	HSR(%)	HWR(%)	HSR(%)	HWR(%)
clean	62.91%	64.93%	59.62%	71.68%	58.13%	62.77%	52.95%	53.16%
$\delta=2/255$	60.34% (-2.57%)	63.14% (-1.79%)	57.37% (-2.25%)	73.14% (+1.46%)	63.35% (+5.22%)	65.87% (+3.10%)	56.07% (+3.12%)	54.76% (+1.60%)
$\delta=5/255$	63.27% (+0.36%)	70.22% (+5.29%)	59.12% (-0.50%)	71.18% (-0.50%)	61.11% (+2.98%)	63.24% (+0.47%)	51.88% (-1.07%)	52.95% (-0.21%)
$\delta=8/255$	60.91% (-2.00%)	66.79% (+1.86%)	57.39% (-2.23%)	69.03% (-2.65%)	58.70% (+0.57%)	62.78% (+0.01%)	51.75% (-1.20%)	52.81% (-0.35%)
attack	68.41% (+5.50%)	70.84% (+5.91%)	65.07% (+5.45%)	75.74% (+4.06%)	64.59% (+6.46%)	67.97% (+5.20%)	57.88% (+4.93%)	59.11% (+5.95%)

shows no significant attack effect in hallucinated content. The comparison highlights the importance of manipulating the inherent behaviors of MLLMs to influence their generation process.

6 Discussions

Alignment of GPT-4 Assisted Evaluation. To assess the alignment and stability of GPT-4-based metrics, we engaged human experts to manually label the hallucinated segments of model responses, and examined the consistency of evaluation results. Detailed analysis is available in Section F of the appendix.

Exploring Mitigation Strategies. In Section 5.5, we observe that the OPERA mitigation [28], though designed to counteract the attention sink behaviors, fails to defend against the adversarial visual inputs. One possible reason is that it only focuses on naturally occurring sink tokens during normal generation processes, and reduces its efficacy in detecting and mitigating the adversarial manipulation of attention. Moreover, our attack achieves consistent success across all tested mitigation methods, underscoring their vulnerabilities in solving natural hallucinations but not considering adversarial inputs adequately. We hope this work inspires future research into more robust defensive strategies for MLLMs, such as those based on adversarial purification with diffusion models [30, 33], the defensive system prompt [70] against hallucination, and new training paradigms to overcome the challenges in instruction-tuning stages.

Adversarial Textual Inputs. Our hallucination attack crafts adversarial visual inputs to achieve high effectiveness and transferability in MLLMs. Considering their multi-modal nature, exploring the perturbation of textual inputs is also valuable. Existing methods for optimizing adversarial texts typically involve gradient-based searches on predefined target responses and modifying discrete tokens to improve attack success [87]. A recent study of jailbreaking attacks against MLLMs suggests decoding adversarial visual inputs within the discrete textual domain [54] to narrow the sampling space. However, the adversarial texts created through these methods generally lack meaningful semantics and are easily detectable. Our approach, on the other hand, optimizes adversar-

ial perturbations by manipulating hidden states and attention mechanisms, which may overcome the current challenges of malicious textual inputs.

Future Works. In our further research, we plan to extend the adversarial efficacy of the proposed attack to a broader range of multi-modal tasks, including visual reasoning, grounding, and multi-turn visual dialogues. Given the rapid adoption of MLLMs in commercial applications, we will also expand our evaluation to more commercial APIs and downstream modules, as evidence of their multi-modal capabilities. Additionally, investigating mitigation strategies through the lens of attention mechanisms and developing defensive approaches are promising directions for future work.

7 Conclusion

This work approaches the prevalent hallucination problem in emerging MLLMs. Through a detailed analysis of the instruction-tuning phase of training, we reveal a critical link between the attention sink phenomenon and hallucinated responses, shedding light on the mechanisms behind erroneous outputs. We propose a novel hallucination attack that induces attention sink behaviors, overcoming the limitations of previous adversarial methods that rely on predefined patterns. Our attack exhibits high transferability, effectively bypassing extensive mitigation strategies and the latest closed-source commercial APIs. We aim to contribute to the safe and reliable development of MLLMs by highlighting current vulnerabilities and inspiring future mitigation strategies.

Acknowledgement

We sincerely appreciate the valuable comments from the shepherd and reviewers that improve the paper’s quality. This work was supported in part by the National Natural Science Foundation of China (62472096, 62172104, 62172105, 62102093, 62102091, 62302101, 62202106). Min Yang is a faculty of the Shanghai Institute of Intelligent Electronics & Systems and Engineering Research Center of Cyber Security Auditing and Monitoring, Ministry of Education, China.

Ethics Considerations

Our work investigates the hallucination problems in current MLLMs and highlights the potential limitations of existing mitigation strategies. To enhance the faithfulness and helpfulness of MLLMs, we have shared our findings and examples with the providers of commercial MLLM APIs analyzed in this study (e.g., OpenAI, Google) via email.

In all the experiments, the model responses only include hallucinated content that is inconsistent with visual inputs, with no harmful or malicious responses generated. All tests conducted with commercial APIs adhere to the platform’s usage guidelines, without any dissemination of hallucinated content or negative impacts on downstream applications.

To facilitate further research on MLLM hallucination, we release our attack algorithms under a restrictive open-sourcing format. The access will be granted only upon request and exclusively for research purposes, mitigating misuse risks. Additionally, the adversarial visual inputs generated during this study will remain confidential to prevent potential harm.

Open Science

We are committed to the principles of open science and have made our source code available upon request for research purposes. The permanent link to the artifacts is <https://zenodo.org/records/14709768>. Researchers are welcome to contact us for access to the artifacts.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.
- [3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.
- [4] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.
- [5] Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The (r) evolution of multimodal large language models: A survey. *arXiv preprint arXiv:2402.12451*, 2024.
- [6] Gongwei Chen, Leyang Shen, Rui Shao, Xiang Deng, and Liqiang Nie. Lion: Empowering multimodal large language model with dual-level visual knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26540–26550, 2024.
- [7] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*, 2024.
- [8] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023.
- [9] Kang Chen, Qingheng Zhang, Chengbao Lian, Yixin Ji, Xuwei Liu, Shuguang Han, Guoqiang Wu, Fei Huang, and Jufeng Chen. Ipl: Leveraging multimodal large language models for intelligent product listing. *arXiv preprint arXiv:2410.16977*, 2024.
- [10] Ke Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm’s referential dialogue magic. *ArXiv*, abs/2306.15195, 2023.
- [11] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023.
- [12] Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Shengyi Qian, Jianing Yang, David F Fouhey, and Joyce Chai. Multi-object hallucination in vision-language models. *arXiv preprint arXiv:2407.06192*, 2024.
- [13] Hao Cheng, Erjia Xiao, Jiahang Cao, Le Yang, Kaidi Xu, Jindong Gu, and Renjing Xu. Typography leads semantic diversifying: Amplifying adversarial transferability across multimodal large language models. *arXiv preprint arXiv:2405.20090*, 2024.
- [14] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023), 2(3):6, 2023.
- [15] Xuanming Cui, Alejandro Aparcedo, Young Kyun Jang, and Ser-Nam Lim. On the robustness of large multimodal models against image adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24625–24634, 2024.

- [16] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv*, abs/2305.06500, 2023.
- [17] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*, 2024.
- [18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.
- [19] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369, 2023.
- [20] Ziliang Gan, Yu Lu, Dong Zhang, Haohan Li, Che Liu, Jian Liu, Ji Liu, Haipang Wu, Chaoyou Fu, Zenglin Xu, et al. Mme-finance: A multimodal finance benchmark for expert-level understanding and reasoning. *arXiv preprint arXiv:2411.03314*, 2024.
- [21] Kuofeng Gao, Yang Bai, Jiawang Bai, Yong Yang, and Shu-Tao Xia. Adversarial robustness for visual grounding of multimodal large language models. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*, 2024.
- [22] Kuofeng Gao, Yang Bai, Jindong Gu, Shu-Tao Xia, Philip Torr, Zhifeng Li, and Wei Liu. Inducing high energy-latency of large vision-language models with verbose images. In *The Twelfth International Conference on Learning Representations*, 2024.
- [23] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*, 2023.
- [24] Yichen Gong, DeLong Ran, Jinyuan Liu, Conglei Wang, Tianshuo Cong, Anyu Wang, Sisi Duan, and Xiaoyun Wang. Figstep: Jailbreaking large vision-language models via typographic visual prompts. *arXiv preprint arXiv:2311.05608*, 2023.
- [25] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024.
- [26] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, 2021.
- [27] Yu-Chih-Tuan Hu, Bo-Han Kung, Daniel Stanley Tan, Jun-Cheng Chen, Kai-Lung Hua, and Wen-Huang Cheng. Naturalistic physical adversarial patch for object detectors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7848–7857, 2021.
- [28] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427, 2024.
- [29] Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27036–27046, 2024.
- [30] Mintong Kang, Dawn Song, and Bo Li. Diffattack: Evasion attacks against diffusion-based adversarial purification. *Advances in Neural Information Processing Systems*, 36, 2024.
- [31] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowd-sourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.
- [32] Gyeong-Geon Lee, Ehsan Latif, Lehong Shi, and Xiaoming Zhai. Gemini pro defeated by gpt-4v: Evidence from education. *arXiv preprint arXiv:2401.08660*, 2023.
- [33] Minjong Lee and Dongwoo Kim. Robust evaluation of diffusion-based adversarial purification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 134–144, 2023.
- [34] Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. Volcano: mitigating multimodal hallucination through self-feedback guided revision. *arXiv preprint arXiv:2311.07362*, 2023.
- [35] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882, 2024.

- [36] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024.
- [37] Xiaoqi Li, Mingxu Zhang, Yiran Geng, Haoran Geng, Yuxing Long, Yan Shen, Renrui Zhang, Jiaming Liu, and Hao Dong. Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18061–18070, 2024.
- [38] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.
- [39] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26763–26773, 2024.
- [40] Chenwei Lin, Hanjia Lyu, Jiebo Luo, and Xian Xu. Harnessing gpt-4v (ision) for insurance: A preliminary exploration. *arXiv preprint arXiv:2404.09690*, 2024.
- [41] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 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, 2014.
- [42] Zheng Lin, Zhenxing Niu, Zhibin Wang, and Yinghui Xu. Interpreting and mitigating hallucination in mllms through multi-agent debate. *arXiv preprint arXiv:2407.20505*, 2024.
- [43] Bo Liu, Ke Zou, Liming Zhan, Zexin Lu, Xiaoyu Dong, Yidi Chen, Chengqiang Xie, Jiannong Cao, Xiao-Ming Wu, and Huazhu Fu. Gemex: A large-scale, groundable, and explainable medical vqa benchmark for chest x-ray diagnosis. *arXiv preprint arXiv:2411.16778*, 2024.
- [44] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2023.
- [45] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.
- [46] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.
- [47] Jiazhen Liu, Yuhan Fu, Ruobing Xie, Runquan Xie, Xingwu Sun, Fengzong Lian, Zhanhui Kang, and Xirong Li. Phd: A prompted visual hallucination evaluation dataset. *arXiv preprint arXiv:2403.11116*, 2024.
- [48] Xin Liu, Yichen Zhu, Yunshi Lan, Chao Yang, and Yu Qiao. Query-relevant images jailbreak large multimodal models. *arXiv preprint arXiv:2311.17600*, 2023.
- [49] Yangzhou Liu, Yue Cao, Zhangwei Gao, Weiyun Wang, Zhe Chen, Wenhui Wang, Hao Tian, Lewei Lu, Xizhou Zhu, Tong Lu, et al. Mminstruct: A high-quality multimodal instruction tuning dataset with extensive diversity. *arXiv preprint arXiv:2407.15838*, 2024.
- [50] Weidi Luo, Siyuan Ma, Xiaogeng Liu, Xiaoyu Guo, and Chaowei Xiao. Jailbreakv-28k: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks. *arXiv preprint arXiv:2404.03027*, 2024.
- [51] Siyuan Ma, Weidi Luo, Yu Wang, Xiaogeng Liu, Muhao Chen, Bo Li, and Chaowei Xiao. Visual-roleplay: Universal jailbreak attack on multimodal large language models via role-playing image character. *arXiv preprint arXiv:2405.20773*, 2024.
- [52] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019.
- [53] Thao Nguyen, Haotian Liu, Yuheng Li, Mu Cai, Utkarsh Ojha, and Yong Jae Lee. Yo’llava: Your personalized language and vision assistant. *arXiv preprint arXiv:2406.09400*, 2024.
- [54] Zhenxing Niu, Haodong Ren, Xinbo Gao, Gang Hua, and Rong Jin. Jailbreaking attack against multimodal large language model. *arXiv preprint arXiv:2402.02309*, 2024.
- [55] OpenAI. Gpt-4v(ision) technical work and authors. 2023.
- [56] OpenAI. Gemini flash. 2024.
- [57] OpenAI. Gpt-4o mini: advancing cost-efficient intelligence. 2024.
- [58] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shao-han Huang, Shuming Ma, Qixiang Ye, and Furu Wei. Grounding multimodal large language models to the world. In *The Twelfth International Conference on Learning Representations*, 2024.

- [59] Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21527–21536, 2024.
- [60] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [61] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, 2018.
- [62] Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. In *The Twelfth International Conference on Learning Representations*, 2023.
- [63] Seungwoo Son, Wonpyo Park, Woohyun Han, Kyuyeun Kim, and Jaeho Lee. Prefixing attention sinks can mitigate activation outliers for large language model quantization. *arXiv preprint arXiv:2406.12016*, 2024.
- [64] Guohao Sun, Can Qin, Huazhu Fu, Linwei Wang, and Zhiqiang Tao. Self-training large language and vision assistant for medical question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20052–20060, 2024.
- [65] Zhiyu Tan, Mengping Yang, Luozheng Qin, Hao Yang, Ye Qian, Qiang Zhou, Cheng Zhang, and Hao Li. An empirical study and analysis of text-to-image generation using large language model-powered textual representation. *arXiv preprint arXiv:2405.12914*, 2024.
- [66] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [67] Ashish Vaswani. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.
- [68] Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.
- [69] Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. Label words are anchors: An information flow perspective for understanding in-context learning. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.
- [70] Yu Wang, Xiaogeng Liu, Yu Li, Muhao Chen, and Chaowei Xiao. Adashield: Safeguarding multi-modal large language models from structure-based attack via adaptive shield prompting. *arXiv preprint arXiv:2403.09513*, 2024.
- [71] Chaoyi Wu, Jiayu Lei, Qiaoyu Zheng, Weike Zhao, Weixiong Lin, Xiaoman Zhang, Xiao Zhou, Ziheng Zhao, Ya Zhang, Yanfeng Wang, et al. Can gpt-4v (ision) serve medical applications? case studies on gpt-4v for multimodal medical diagnosis. *arXiv preprint arXiv:2310.09909*, 2023.
- [72] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*, 2024.
- [73] Qianqian Xie, Dong Li, Mengxi Xiao, Zihao Jiang, Ruoyu Xiang, Xiao Zhang, Zhengyu Chen, Yueru He, Weiguang Han, Yuzhe Yang, et al. Open-finllms: Open multimodal large language models for financial applications. *arXiv preprint arXiv:2408.11878*, 2024.
- [74] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*, 2024.
- [75] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, et al. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv preprint arXiv:2310.05126*, 2023.
- [76] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.
- [77] Zonghao Ying, Aishan Liu, Xianglong Liu, and Dacheng Tao. Unveiling the safety of gpt-4o: An empirical study using jailbreak attacks. *arXiv preprint arXiv:2406.06302*, 2024.
- [78] Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12944–12953, 2024.
- [79] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rllhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816, 2024.

- [80] Zhongzhi Yu, Zheng Wang, Yonggan Fu, Huihong Shi, Khalid Shaikh, and Yingyan Celine Lin. Unveiling and harnessing hidden attention sinks: Enhancing large language models without training through attention calibration. *arXiv preprint arXiv:2406.15765*, 2024.
- [81] Zihao Yue, Liang Zhang, and Qin Jin. Less is more: Mitigating multimodal hallucination from an eos decision perspective. *arXiv preprint arXiv:2402.14545*, 2024.
- [82] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. *Advances in Neural Information Processing Systems*, 36, 2024.
- [83] Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*, 2023.
- [84] Juexiao Zhou, Xiaonan He, Liyuan Sun, Jiannan Xu, Xiuying Chen, Yuetan Chu, Longxi Zhou, Xingyu Liao, Bin Zhang, Shawn Afvari, et al. Pre-trained multimodal large language model enhances dermatological diagnosis using skingpt-4. *Nature Communications*, 15(1):5649, 2024.
- [85] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *The Twelfth International Conference on Learning Representations*, 2024.
- [86] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.
- [87] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

A Details of Implementation

During the hallucination attack, we set the total steps S_{\max} of adversarial optimization as 30, with a learning rate γ of 5 to update the adversarial perturbations. The attack budget ϵ is varied across 2/255, 5/255, and 8/255 to modulate the perturbation magnitude. Following an ablation study, we set the hyper-parameter $\alpha = 1$ to optimize attack effectiveness. When retrieving hidden states from intermediate levels, we select the second-to-last layer for InstructBLIP and LLaVA-1.5, and the third-to-last layer for MiniGPT-4 and Shikra. During model response generation with beam search decoding, we configured the beam width N_{beam} to 3.

B Results of Human-Evaluated Hallucination Types

To further analyze the adversarial effects of our attack, we engaged human experts to classify the types of hallucinations. The evaluation focused on model responses from the image captioning task across 6 MLLMs, with results detailed in Fig. 9. Open-source MLLMs demonstrate a higher propensity for generating fabricated content when exposed to adversarial visual inputs, whereas commercial APIs more frequently misidentify attributes or relationships between objects.

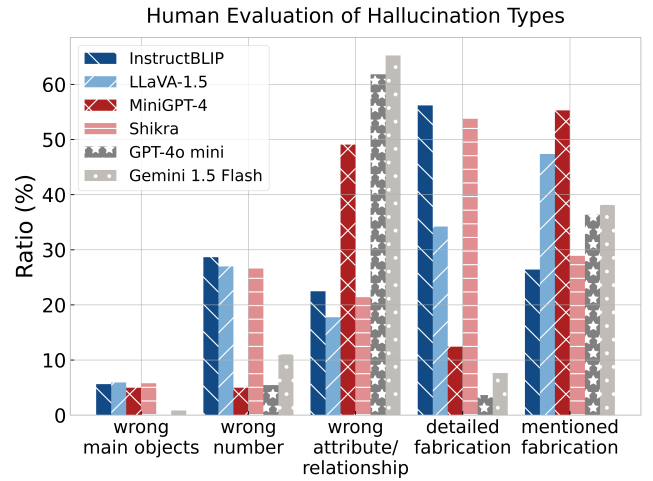


Figure 9: Results of the human evaluation on hallucination types. The five types of hallucinations are: identifying *wrong main objects*, *wrong number* of objects, *wrong attributes or relationships* to objects, generating *detailed fabrication* content, and producing only *mentioned fabrication* content.

C More Results of Black-box Transferability

The adversarial QA accuracy results on black-box commercial APIs are illustrated in Fig. 10.

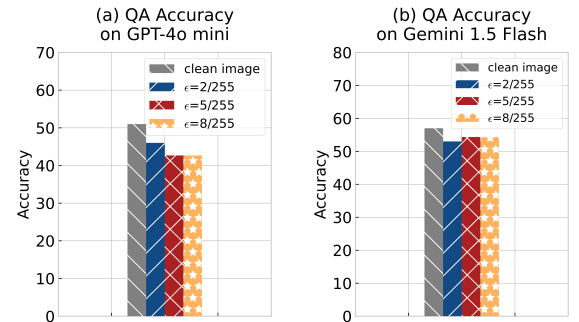


Figure 10: Results of QA Accuracy on commercial APIs: (a) GPT-4o mini released by OpenAI and (b) Gemini 1.5 Flash launched by Google. A lower accuracy indicates more hallucinations in MLLM responses.

Table 8: Comparison of human-evaluated and GPT-4 assisted hallucination metrics on 4 sets of model responses. (\dagger) denotes the evaluation results of human experts, while (\circ) denotes the evaluation results of GPT-4. The human-evaluated results are averaged between two experts. The number in brackets indicates the hallucination improvement compared to the clean image, with human-evaluated results in **red color** and GPT-4-evaluated results in **indigo color**. Adversarial results are marked in bold.

	Input	HSPI	HWPI	HSR(%)	HWR(%)
MiniGPT-4	clean image \dagger	2.53	39.84	49.32%	51.21%
	clean image \circ	2.92	44.61	54.42%	56.02%
	$\epsilon=8/255^\dagger$	2.79 (+0.26)	43.04 (+3.20)	52.10% (+2.78%)	54.00% (+2.79%)
	$\epsilon=8/255^\circ$	2.92	46.00 (+1.39)	57.38% (+2.96%)	59.42% (+3.40%)
LLaVA-1.5	clean image \dagger	2.16	41.31	42.44%	45.76%
	clean image \circ	2.27	42.77	45.17%	48.11%
	$\epsilon=8/255^\dagger$	2.41 (+0.25)	46.62 (+5.31)	47.58% (+5.14%)	51.36% (+5.60%)
	$\epsilon=8/255^\circ$	2.67 (+0.40)	50.23 (+7.46)	53.04% (+7.87%)	55.78% (+7.67%)

D Results of Perplexity-based Model Response Quality

In Section 5.4, we present the results of the GPT-4-assisted evaluation of model response quality. The Perplexity-based quality evaluation results are presented in Fig. 11, calculated with the pre-trained GPT-2 model.

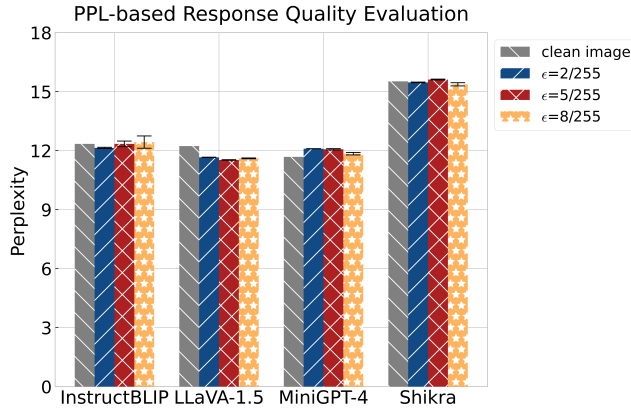


Figure 11: Results of Perplexity-based response quality evaluation, covering both white-box and black-box attack scenarios. A lower perplexity reflects better MLLM response quality.

E Results of Adaptive Mitigation

Considering the adaptive mitigation strategy of detecting and early-stopping before sink tokens, we present the length, quality, and hallucination metrics of MLLMs responses under the mitigation strategy in Tab. 9.

F Results of Alignment Study

To assess the alignment between GPT-4-assisted hallucination evaluation and human judgments, we engaged human experts to manually label the hallucinated segments of model

responses. We compare the HSPI/HWPI/HSR/HWR metrics derived from GPT-4-based evaluations with those from human assessments. The detailed results are presented in Tab. 8. Notably, the human-evaluated hallucination metrics are slightly lower than those obtained from GPT-4, likely due to the limited scope of descriptions and the factual information annotated in the HalluBench dataset. Additionally, the results for adversarial visual inputs show comparable improvements, underscoring the reliability and effectiveness of GPT-4-assisted evaluation in this study.

Table 9: Results of the length (WPI), hallucination (HWR), and response quality under the early-stopping mitigation strategy. clean* refers to the results of clean images on MLLM without any mitigation strategy applied, while clean \circ indicates the results on MLLM with adaptive mitigation. The best adversarial results are marked in bold.

Target Model	Input	WPI	HWR(%)	Quality
InstructBLIP	clean*	102.89	77.27%	8.58
	clean \circ	46.51	52.56%	5.55
	$\epsilon=2/255$	48.57	63.08% (+10.52%)	5.45
	$\epsilon=5/255$	45.31	65.24% (+12.68%)	5.57
	$\epsilon=8/255$	46.31	69.99% (+17.43%)	5.28
MiniGPT-4	clean*	79.70	56.02%	8.81
	clean \circ	51.05	53.67%	6.39
	$\epsilon=2/255$	50.81	58.93% (+5.26%)	6.36
	$\epsilon=5/255$	52.47	56.64% (+2.97%)	6.49
	$\epsilon=8/255$	51.88	59.87% (+6.20%)	5.69