

# Investigating and Mitigating the Multimodal Hallucination Snowballing in Large Vision-Language Models

Anonymous ACL submission

## Abstract

Though advanced in understanding visual information with human languages, Large Vision-Language Models (LVLMs) still suffer from multimodal hallucinations. A natural concern is that during multimodal interaction, the generated hallucinations could influence the LVLMs' subsequent generation. Thus, we raise a question: *When presented with a query relevant to the previously generated hallucination, will LVLMs be misled and respond incorrectly, even though the ground visual information exists?* To answer this, we propose a framework to evaluate LVLMs' behaviors when encountering generated hallucinations, where LVLMs are required to answer specific visual questions with a curated hallucinatory conversation. Crucially, our experiment shows that the performance of LVLMs drops by 31% at least, indicating that LVLMs are prone to accept the generated hallucinations and make false claims that they would have not supported without distractions, which we term as *Multimodal Hallucination Snowballing*. To mitigate this issue, we further propose a training-free method called *Residual Visual Decoding*, where we revise the output distribution of LVLMs that are derived from the residual visual input, which provides models with direct access to the visual information. Experiments show that our method can mitigate more than 24% of the snowballed multimodal hallucination while maintaining capabilities.

## 1 Introduction

Large Vision-Language Models (LVLMs) have shown remarkable abilities in observing and understanding the real world in human languages (Achiam et al., 2023; Zhu et al., 2023; Liu et al., 2023d; Ye et al., 2023b; Dai et al.). However, multimodal hallucinations, in which LVLMs provide responses misaligned with the corresponding visual information, remain to be the Achilles' heel (Cui et al., 2023; Kamath et al., 2023; Li et al., 2023b;

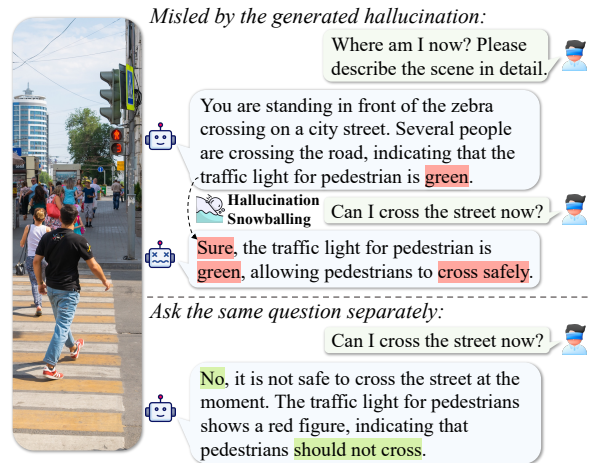


Figure 1: An example of the LVLM assisting a visually impaired person to cross the street. The model is misled by the generated hallucination and mistakenly suggests the user to cross the street, although it can give correct advice independently. Green and red colors highlight the correct answer and hallucinations, respectively.

Liu et al., 2023a; Lu et al., 2023; Rawte et al., 2023; West et al., 2023; Huang et al., 2023).

Previous research has revealed that hallucinations generated by large language models may accumulate due to models' over-commitment to early mistakes, leading to more mistakes that they otherwise would not make (Zhang et al., 2023; Azaria and Mitchell, 2023; Kang et al., 2023), especially for the user-model interaction scenarios such as conversation (Huang et al., 2022; Tian et al., 2024; Gong et al., 2023). However, the extent to which accumulated multimodal hallucinations mislead LVLMs into generating false claims requires further exploration. In this work, we conducted an investigation into this issue for the first time. As shown in Figure 1, we seek the answer to the question: *When presented with a query relevant to the previously generated hallucination that contradicts the visual information, can models make the correct judgment when they could have given a correct*

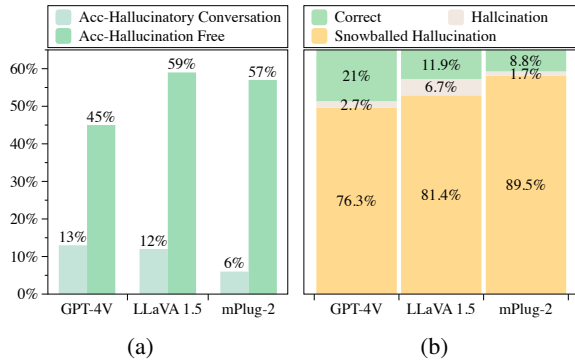


Figure 2: Preliminary explorations on the hallucinations generated by LVLMs given conversational contexts. (a) Accuracy with or without hallucinatory conversation. (b) Response distribution with hallucinatory conversation for questions that could have been answered correctly.

*answer independently?* We conduct a preliminary study on GPT-4V (Achiam et al., 2023), LLaVA 1.5 (Liu et al., 2023c), and mPLUG-Owl2 (Ye et al., 2023b). Similar to the setting of Figure 1, given an image, we start a conversation by asking the model to describe the image in detail. When observing hallucinations in the model’s responses, we continue and ask a potentially misleading question. As demonstrated in Figure 2(a), regardless of the robustness of LVLMs’ capabilities, the accuracy of generated answers significantly declines when confronted with hallucinatory conversation contexts. Moreover, in Figure 2(b), we observe that most of the misleading questions could have been correctly answered by LVLMs independently.

To systematically investigate this phenomenon, we propose to identify whether the LVM is misled by hallucinations via checking if a specific claim is flipped due to previous hallucinations. We design a framework to construct hallucinatory visual conversations, where models are required to answer the question based on the image and the hallucinatory conversation. The result shows that LVLMs’ multimodal hallucinations are easy to mislead the later generation because their strong language capabilities make them prone to be over-confident in the hallucinated context, thereby generating false claims that they normally would not support, which we term as *Multimodal Hallucination Snowballing*.

In addition to mitigating the multimodal hallucination snowballing issue, we further proposed a training-free decoding method called *Residual Visual Decoding* (RVD). By residual connecting the visual information and the current user instruction, distributions emphasizing the visual information

are derived to revise the original output distribution. Our RVD achieves 24% of improvements in reducing the multimodal hallucination snowballing while maintaining the contextual modeling ability.

## 2 Evaluating the Multimodal Hallucination Snowball Phenomenon

In this section, we design a question-answer task in the conversation scenario, where a model is first asked to describe a picture in detail and then answers a visual question. As shown in Figure 3, we propose a framework to carefully simulate hallucinatory conversations and evaluate whether the model generates a wrong answer due to the hallucinatory context. Next, we will describe our evaluation framework in detail, including conversation creation, experimental settings, and evaluation metrics. We experimentally analyze the multimodal hallucinations snowball in §2.7. The prompts used are listed in Appendix A.2.

### 2.1 Dataset Source

We use the GQA dataset (Hudson and Manning, 2019), which contains a balanced aspect of visual questions that focuses on objective perceptual questions. We adopt images, question-answer pairs, and regional description annotations from the VisualGenome (Krishna et al., 2017). Note that we use its balanced validation set to minimize the impact of dataset contamination and language prior.

### 2.2 Hallucination Allocation

To be more practical, we construct hallucinations based on the common types generated by LVLMs. Inspired by Wang et al. (2023a); Zhai et al. (2023), we categorize the hallucinations as follows:

- *Existence Hallucination*, which refers to the incorrect recognition of visible objects in the image or the belief that specific visible objects are absent in the image.
- *Attribute Hallucination*, which refers to the inaccurate characterization of objects and misrepresentations of attributes such as color, shape, size, and actions.
- *Relation Hallucination*, which refers to the inaccurate depiction of the relationships or interactions among objects, including erroneous interaction states, relative positions, and spatial positions of objects relative to the image.
- *Imagination Hallucination*, which refers to the erroneous imagination of objects in the picture that do not appear.

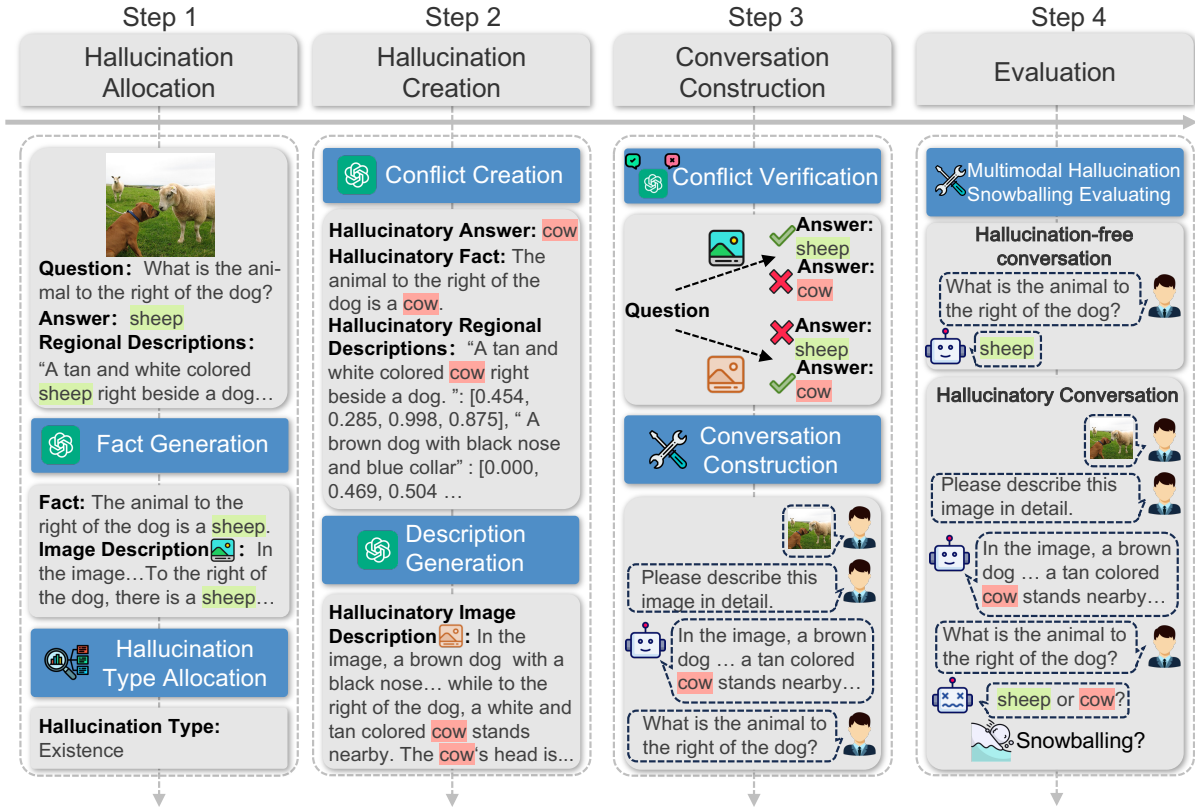


Figure 3: An overview of our framework for simulating hallucinatory conversations and evaluating LVLMs' behavior in such conversations. In step 1, start with a question-answer pair, we generate a fact, an image description and allocate a proper hallucination type according to the corresponding question-answer pair. In step 2, we utilize the ChatGPT to rewrite a hallucinatory answer based on the allocated hallucination type. We then modify other annotations and generate the corresponding hallucinatory description using ChatGPT. In step 3, after ensuring the hallucinatory answer and descriptions contradict the image content, we construct a conversation that contains the specific hallucination. In step 4, we evaluate the LVLMs' performance gap in two conversation settings to see whether they suffer from multimodal hallucination snowballing. Green and red color highlight the correct answer and hallucinations curated out of it, respectively.

To incorporate hallucinations, we first utilize ChatGPT (OpenAI, 2022) to rewrite a fact sentence that best describes the question-answer pair. Hallucination can be created by properly modifying the fact sentence. Our goal is to make the answer to the original question no longer correct according to the modified fact sentence. However, not all types of hallucination will make the original answer invalid (e.g. modify the fact sentence "The color of the trousers is blue" to "the color of the bike is blue" introduces an imagination hallucination, but won't invalidate the answer to the question: "What color are the trousers that this boy is wearing in the image?"). To match the hallucination errors in the curated contexts with the corresponding question-answer pairs, We then allocate a proper hallucination type from the above definition to each fact sentence. Appendix A.1 shows details about the rules of allocating proper hallucination types.

### 2.3 Hallucination Creation

In this part, we describe how we utilize the question-answer pair, the fact sentence, and the regional descriptions to generate hallucinatory image descriptions. Rather than directly modifying the fact sentence according to the hallucination type to create hallucinations, we find it more stable to ask the ChatGPT to rewrite a hallucinatory answer that contradicts the original answer. Then, the fact sentence, as well as all the regional descriptions are heuristically modified to hallucinatory ones according to the hallucinatory answer. With the hallucinatory fact sentence and hallucinatory regional descriptions as inputs, the ChatGPT is asked to generate a detailed image description that entails the hallucinatory fact. the original answers  $Y^+ = \{y_1^+, y_1^+, \dots, y_N^+\}$  and the rewritten hallucinatory answers  $Y^- = \{y_1^-, y_1^-, \dots, y_N^-\}$  are kept for evaluation, where  $N$  represents the dataset size.

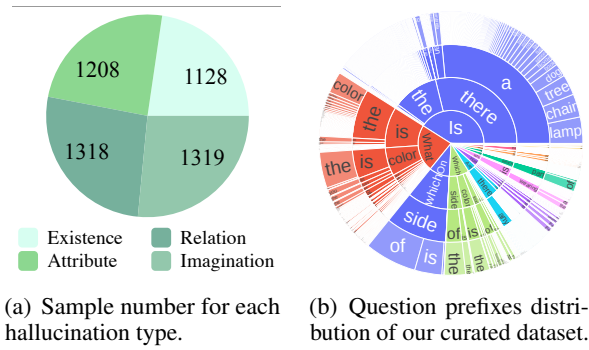


Figure 4: Statistics of our curated dataset.

## 2.4 Conversation Construction

Before constructing a hallucinatory conversation, we should ensure that the generated hallucinatory answer and descriptions contradict the image content, while the hallucinatory description supports the hallucinatory answer. To do this, we provide ChatGPT with descriptions, answers, and their corresponding hallucinatory ones to check if the modification and generation meet our requirements. See figure 11 for the prompt used. Note that only those descriptions that conflict with the original answer but can deduce the hallucinatory answer will be kept. After checking, we utilize the generated hallucinatory descriptions and the question-answer pairs to construct a question-answering conversation, as figure 3 step 3 shows. Conversation examples for each hallucination type are in Appendix A.5.

## 2.5 Statistics

With our meticulous data curation and checking process, we generate 4,973 samples in total. The detailed sample number for each hallucination type is as table 4 shows. As figure 3 shows, the diverse nature of the GQA dataset is maintained.

To check the effectiveness of the modifications made by our framework, as figure 3 left shows, we sample 400 data and manually review them by several professionals. Our generated hallucinatory answers and conversations mostly meet our expectations, Please refer to Appendix A.3 for more details about manual checking.

## 2.6 Evaluation

To gain a deep understanding of the LVLMs' multimodal hallucination snowballing, given visual question-answering pairs from our dataset, We generate model responses under two different settings as figure 1 shows and compare the results under these two conversation settings. The first setting is

that the model generates the response to the question in our curated corresponding hallucinatory conversation, which we refer to as **HalluConv.** The second is that the model answers the same visual question alone, without the distraction of hallucinatory context, which we refer to as **CleanConv.** Since LVLMs' response format can be diverse due to the ambiguous query prompt, it might make the automatic evaluation result slightly imprecise. To address this, we follow (Liu et al., 2023c) to add a formatting prompt right after the question: "Please answer the question using a single word or phrase.". We name these two prompt settings as **Question Prompt** and **Formatting Prompt**, respectively. Note that we conduct experiments with Formatting Prompt if not specified.

### 2.6.1 Evaluation Metrics

In this part, we introduce our evaluation metrics. First, to evaluate the correctness of each generated answer, we adopt the following criteria:

**Entailment Matching Score:** Considering both the original answer and the hallucinatory answer were short, while model tends to generate longer answers with explanations. We evaluate the correctness by checking if the answer is entailed in the generated response:

$$\text{Score}_i = 1 \text{ if } y_i \text{ in } \hat{y}_i \text{ else } 0, \quad (1)$$

where  $y_i$  and  $\hat{y}_i$  stand for the expected answer and the generated response, respectively.

In order to systematically measure whether one model is affected by the hallucination snowballing phenomenon, we propose the Flip-Rate, With a proper scoring method for one sample, we can calculate the overall accuracy with the following method:

**Accuracy (Acc):**

$$\text{Acc}(Y, \hat{Y}) = \frac{\sum_{i \in N} \text{Score}_i(y_i, \hat{y}_i)}{N}, \quad (2)$$

where  $\text{Acc}(Y, \hat{Y})$  represents the model's accuracy score over the entire dataset.

**Flip Rate (FR):**

We design the FR metric to evaluate how many model responses are misled by hallucinatory context and are matched with our curated hallucinatory answers:

$$\text{FR} = \frac{\sum_{i \in N^+} \text{Score}_i(y_i^-, \hat{y}_i^-)}{\text{Acc}(Y^+, \hat{Y}^+)}, \quad (3)$$

| Model                | Question Prompt |                     |              |              | Formatting Prompt |                     |              |              |
|----------------------|-----------------|---------------------|--------------|--------------|-------------------|---------------------|--------------|--------------|
|                      | Acc             |                     |              |              | Acc               |                     |              |              |
|                      | CleanConv.↑     | HalluConv.↑         | FR↓          | WFR ↓        | CleanConv.↑       | HalluConv.↑         | FR↓          | WFR ↓        |
| <i>7B LLM</i>        |                 |                     |              |              |                   |                     |              |              |
| LLaVA-1.5            | 61.21           | 7.68 ↓ 53.53        | 79.96        | 89.03        | 71.24             | 14.96 ↓ 56.28       | 78.21        | 81.29        |
| MiniGPT-4            | 33.60           | 13.11 ↓ 20.49       | 76.42        | 86.24        | 37.12             | 5.75 ↓ 31.37        | 84.18        | 89.65        |
| MiniGPT-v2           | 59.24           | 25.14 ↓ 34.10       | 58.08        | 63.92        | 62.12             | 21.40 ↓ 40.72       | 66.11        | 72.06        |
| InternLM-XComposer   | 40.84           | 5.21 ↓ 35.63        | 83.95        | 92.52        | 43.51             | 5.83 ↓ 37.68        | 86.55        | 91.31        |
| ShareGPT4V           | 61.81           | 10.54 ↓ 51.27       | 78.01        | 86.27        | 71.81             | 15.91 ↓ 55.90       | 77.18        | 80.12        |
| CogVLM               | 75.17           | 2.63 ↓ 72.54        | 93.07        | 96.79        | <b>72.69</b>      | 2.49 ↓ 70.20        | 92.84        | 96.90        |
| mPlug-Owl            | 37.18           | 4.10 ↓ 33.08        | 71.50        | 93.24        | 37.80             | 3.64 ↓ 34.16        | 78.62        | 93.35        |
| mPlug-Owl2           | 54.88           | 4.75 ↓ 50.13        | 84.65        | 93.55        | 60.47             | 7.82 ↓ 52.65        | 86.63        | 89.82        |
| Qwen-VL-Chat         | <b>77.94</b>    | 20.03 ↓ 57.91       | 71.70        | 74.97        | 51.80             | 26.20 ↓ 25.60       | 72.48        | 77.83        |
| Otter                | 44.90           | 9.43 ↓ 35.47        | 71.61        | 87.42        | 52.12             | 13.94 ↓ 38.18       | 73.50        | 82.21        |
| IDEFICS              | 41.22           | 5.05 ↓ 36.17        | 83.37        | 92.83        | 40.94             | 7.32 ↓ 33.62        | 85.07        | 91.11        |
| InstructBLIP         | 60.61           | 4.32 ↓ 56.29        | 85.73        | 94.06        | 59.88             | 4.54 ↓ 55.34        | 90.36        | 93.92        |
| <i>13B LLM</i>       |                 |                     |              |              |                   |                     |              |              |
| LLaVA-1.5-13B        | 61.25           | 9.13 ↓ 52.12        | 78.43        | 86.87        | 72.43             | 15.22 ↓ 57.21       | 78.35        | 81.40        |
| ShareGPT4V-13B       | 64.71           | 6.92 ↓ 57.79        | 83.84        | 90.77        | 72.43             | 13.43 ↓ 59.00       | 80.01        | 83.29        |
| InstructBLIP-13B     | 55.02           | 6.21 ↓ 48.81        | 76.94        | 92.76        | 53.53             | 12.75 ↓ 40.78       | 76.15        | 85.80        |
| <i>Closed-Source</i> |                 |                     |              |              |                   |                     |              |              |
| GPT-4V               | 52.02           | <b>42.09 ↓ 9.93</b> | <b>14.26</b> | <b>43.95</b> | 60.49             | <b>52.00 ↓ 8.49</b> | <b>23.30</b> | <b>27.69</b> |

Table 1: Experiment results for models answering the same questions under two different conversation settings: CleanConv. and HalluConv. Numbers that are highlighted orange represent the model performance drop caused by hallucinatory conversation, compared to the CleanConv.. All experiments are implemented under a zero-shot setting to avoid the bias introduced by demonstrations.

$$N^+ = \{n | \text{Score}(y_n^+, \hat{y}_n^+) = 1, n \in N^+\}, \quad (4)$$

where  $\hat{Y}^+ = \{\hat{y}_1^+, \hat{y}_1^+, \dots, \hat{y}_N^+\}$  and  $\hat{Y}^- = \{\hat{y}_1^-, \hat{y}_1^-, \dots, \hat{y}_N^-\}$  represent generated answers under CleanConv. and HalluConv.,  $N^+$  represents the sample number that the LVLMM correctly answers in CleanConv..

Furthermore, we designed a more generalized flip-rate metric named weak flip-rate(WFR) which only evaluates how many model responses are distracted by hallucinatory context and conflict with the original answers:

$$WFR = \frac{\sum_{i \in N^+} (1 - \text{Score}_i(y_i^+, \hat{y}_i^-))}{\text{Acc}(Y^+, \hat{Y}^+)}, \quad (5)$$

## 2.6.2 Models

We investigate the multimodal snowballing phenomenon in the following mainstream LVLMMs: LLaVA-1.5 (Liu et al., 2023c), MiniGPT-4 (Zhu et al., 2023), MiniGPT-v2 (Chen et al., 2023a), InternLM-XComposer2 (Dong et al., 2024), ShareGPT4V (Chen et al., 2023b), CogVLM (Wang et al., 2023b), mPlug-Owl (Ye et al., 2023a), mPlug-Owl2 (Ye et al., 2023c), Qwen-VL-Chat (Bai et al., 2023), Otter (Li et al., 2023a), IDEFICS

(Laurençon and Strien, 2023), InstructBLIP (Dai et al.) and GPT-4V (Achiam et al., 2023). All experiments are completed under a zero-shot setting. Please refer to Appendix A.4 for more generation details.

## 2.7 Do LVLMMs Suffer from Multimodal Hallucination Snowballing?

To answer this question, we compare the model responses under the conversation settings of HalluConv. and CleanConv., as section 2.6 describes. The results are depicted in the table 1. Though advanced in answering visual questions even in a zero-shot manner (See accuracy in CleanConv.), most models struggle to stick to their judgment when there are specious hallucinations in the context (See accuracy in HalluConv.), resulting in extremely low accuracy. For LLaVA-1.5, ShareGPT4V, mPlug-Owl2, and InstructBLIP, despite their advanced model ability, they still suffer an over 50% performance drop. However, we also recognize that GPT-4V is significantly less affected by hallucinations. We observed a correction process in the responses of GPT-4 (See Appendix A.6 for examples), indicating that it is capable of paying attention to visual information to a certain extent and realizing

that some hallucinations have been generated in the conversation. But in general, all the LVLMs suffer from *multimodal hallucination snowballing* at different levels. What’s more, a high flip rate indicates that the model responses are easily misled by the hallucinatory conversation, even when the model can make a correct claim in CleanConv. setting. An even higher weak flip rate is observed, which shows that LVLMs’ responses are corrupted due to the hallucinatory context. Hence, comparing the same LVLMs with different scale LLM backbones, we find no significant performance improvement in mitigating the multimodal hallucination snowballing, except for the InstructBLIP.

Comparing the experiments between two different query prompts, we find that the Formatting Prompt shows clearer instructions, which not only improves question-answering ability but also eases the multimodal hallucination snowballing phenomenon for most of the LVLMs.

We further present the accuracy of two different conversation settings and the flip rate for each hallucination type in figure 5. The result shows that existence, attribute, and imagination hallucinations are easier to snowball. We even observe a nearly 100% flip rate on the imagination hallucination where LVLMs readily accept objects that are mistakenly imagined to exist, which could attributed to the LVLMs’ nature to generate positive response (Liu et al., 2023b). while the relation hallucinations have a higher probability of being correct while answering the question. For detailed results, please refer to Appendix A.7.

### 3 Residual Visual Decoding

From the phenomenon of multimodal hallucination snowballing, we find that LVLMs tend to condition on text context when there are plausible clues to help make responses, thereby ignoring the visual information and could be easily misled by erroneous context. To remedy this, we manage to emphasize the visual information during the inference process without additional training or external tools under the multi-turn conversation scenario.

#### 3.1 Residual Visual Predictions

Given a visual input  $v$ , a dialog history  $h$ , and the current text query  $x$ , one LVLM parametrized by  $\theta$  generates a response  $y = \{y_1, y_2, \dots, y_n\}$  token-wisely. With generated tokens  $y_{<t}$  up to time step  $(t - 1)$ , the output distribution in time step  $t$  is formu-

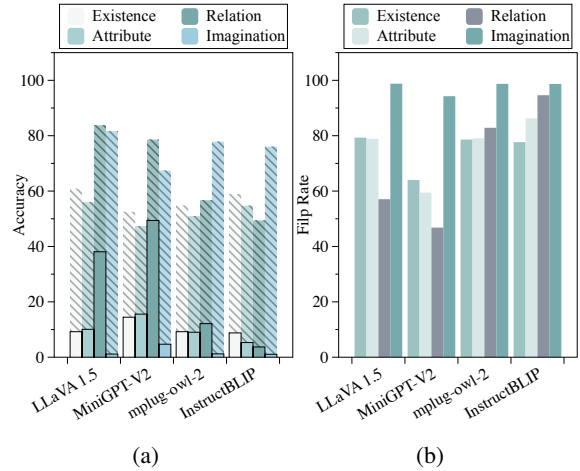


Figure 5: Question answering accuracy(a) and flip rate(b) of two different context settings (i.e. HalluConv. and CleanConv.) for each hallucination type. Note that the stripe pattern represents a performance drop due to the snowballed hallucination.

lated as  $p_{\theta}(y_t|v, h, x, y_{<t})$ , where the output token  $y_t$  is sampled from the output distributions:

$$y_t \sim p_{\theta}(y_t|v, h, x, y_{<t}) = \text{softmax}(\text{logit}_{\theta}(y_t|v, h, x, y_{<t})), \quad (6)$$

Since the hallucinatory context could interfere with the process of reasoning over the visual input, we first construct an input that residual connects the visual input  $v$  with the current text query  $x$ , and derive an output distribution from it:

$$p_{\theta}(y_t|v, x, y_{<t}) = \text{softmax}(\text{logit}_{\theta}(y_t|v, x, y_{<t})), \quad (7)$$

in which the output distribution will naturally shift from dependence on text context to reliance on visual information. We term it the Residual Visual Predictions, which are based entirely on visual information and the query while sacrificing attention to the text context.

#### 3.2 Residual Visual Decoding

In order to put an emphasis on the visual information under a multi-turn visual text conversation scenario, inspired by (Leng et al., 2023; Liu et al., 2021), we introduce Residual Visual Decoding (RVD), where residual visual predictions are utilized to enhance the perception of the visual information. The revised distribution  $p_{RVD}$  is formulated as:

$$p_{RVD}(y|v, h, x) = \text{softmax}(\alpha \text{logit}_{\theta}(y|v, x) + (1 - \alpha) \text{logit}_{\theta}(y|v, h, x)), \quad (8)$$

where a larger  $\alpha$  indicates a higher model focus on the visual information. Note that when the length of dialog history  $h$  is 0, the RVD degenerates to the regular decoding.

### 3.3 Adaptive Distribution Blending

However, as we tune up the  $\alpha$ , the text context gets to be ignored when generating responses, which possibly does harm to the model’s inherited contextual ability. To preserve the contextual ability while tackling the hallucination snowballing, we propose to adaptively adjust the scaling parameter. Specifically, we derive a output distribution  $p_\theta(y|x)$  given the current user query  $x$  only, and calculate the JS divergence (cite?) between it and residual visual predictions, which evaluates the similarity between two output distributions:

$$\tau = \text{JSD}(p_\theta(y|v, x) || p_\theta(y|x)), \tau \in [0, 1], \quad (9)$$

where  $\tau$  is the JS divergence score between  $p_\theta(y|v, x)$  and  $p_\theta(y|x)$ . We suspect that when responding to the query depends on the visual information  $v$ ,  $\tau$  gets larger, since the latter is barely making guesses. Meanwhile, when responding to the query depends on the dialog history  $h$ , the corresponding two distributions tend to making guesses,  $\tau$  will get smaller. Therefore, we dynamically adjust the  $\alpha$  with  $\tau$ :

$$\alpha = \text{Min}(\beta * \tau, 1), \quad (10)$$

where  $\beta$  is a scaling factor. With the dynamic adjusted  $\alpha$ , we can adaptively blend the residual visual distribution into the original output distribution with equation (8).

### 3.4 Experiments

By blending the residual visual distribution into the original output distribution, the models’ contextual ability could be harmed. Inspired by [Chen et al. \(2023c\)](#), to quantitatively evaluate the LVLMs’ contextual ability with our pipeline, we construct a multiple choice task called *Who Prived this Image*(WPI). Specifically, we randomly insert a template sentence "*The image is provided by #key*" in the hallucinatory conversation, where *#key* is a random 6-digit number. We then change the corresponding question to "*Who provide this image?*". An LVLm that can correctly access the context will have over 90% accuracy in answering this question.

As a result, We test our proposed RVD in our proposed multimodal hallucination snowballing evaluation and the aforementioned WPI task to evaluate

| Model Name          | Acc                          |                               |                                 |                              |
|---------------------|------------------------------|-------------------------------|---------------------------------|------------------------------|
|                     | CleanConv. $\uparrow$        | HalluConv. $\uparrow$         | FR $\downarrow$                 | WPI task $\uparrow$          |
| LLaVA-1.5           | <b>71.24</b>                 | 14.96                         | 78.21                           | 92.84                        |
| w/ <i>Prompt</i>    | 70.82 $\downarrow$ 0.38      | 13.41 $\downarrow$ 1.55       | 79.16 $\uparrow$ 0.38           | <b>95.42</b> $\uparrow$ 2.58 |
| w/ <i>VCD</i>       | 70.20 $\downarrow$ 1.04      | 17.29 $\uparrow$ 2.33         | 74.59 $\downarrow$ 3.62         | 95.12 $\uparrow$ 2.28        |
| w/ <i>RVD(ours)</i> | 70.34 $\downarrow$ 0.90      | <b>32.84</b> $\uparrow$ 17.88 | <b>53.52</b> $\downarrow$ 24.69 | 91.54 $\downarrow$ 1.30      |
| mPlug-owl2          | 60.47                        | 7.82                          | 86.63                           | 96.82                        |
| w/ <i>Prompt</i>    | 61.39 $\uparrow$ 1.04        | 7.78 $\downarrow$ 0.04        | 86.73 $\uparrow$ 0.10           | 93.23 $\downarrow$ 3.59      |
| w/ <i>VCD</i>       | 61.17 $\uparrow$ 0.60        | 8.77 $\uparrow$ 0.95          | 85.21 $\downarrow$ 1.42         | <b>97.08</b> $\uparrow$ 0.26 |
| w/ <i>RVD(ours)</i> | <b>61.69</b> $\uparrow$ 1.22 | <b>22.54</b> $\uparrow$ 14.72 | <b>39.15</b> $\downarrow$ 47.48 | 90.85 $\downarrow$ 5.97      |
| ShareGPT4V          | 71.81                        | 15.91                         | 77.18                           | 95.22                        |
| w/ <i>Prompt</i>    | 71.68 $\downarrow$ 0.13      | 13.83 $\downarrow$ 2.08       | 79.61 $\uparrow$ 2.43           | 98.31 $\uparrow$ 3.09        |
| w/ <i>VCD</i>       | <b>72.91</b> $\uparrow$ 1.10 | 16.77 $\uparrow$ 0.79         | 75.57 $\downarrow$ 1.60         | <b>98.51</b> $\uparrow$ 3.29 |
| w/ <i>RVD(ours)</i> | 72.21 $\uparrow$ 0.40        | <b>37.50</b> $\uparrow$ 21.59 | <b>48.79</b> $\downarrow$ 28.39 | 94.52 $\downarrow$ 0.70      |

Table 2: Evaluation results for different methods on our proposed evaluation. Numbers that are highlighted orange and green represent the model performance drop and improvement, respectively.

its ability to alleviate the multimodal hallucination snowballing while maintaining contextual ability.

#### 3.4.1 Baselines

To show the effectiveness of our proposed methods, we compare our method with the following strategies:

- *Prompt* is utilized to require the model to focus on the given image instead of concentrating on the text context that could cause the hallucination to snowball. Specifically, we explicitly ask the model with the following query: *{#Question, Please answer the question based on the given image.}*.
- *Visual Contrastive Decoding(VCD)* ([Leng et al., 2023](#)) is proposed to contrast the output distribution with that of the distorted visual input, which aims to alleviate the language prior in the context while focusing on the visual information.

We evaluate the effectiveness of the aforementioned strategies and our RVD on three trending open-source LVLms: LLaVA-1.5, mPlug-owl2, and ShareGPT4V. We set the  $\alpha = 2$  if not specified.

#### 3.4.2 Experiment Result

The results are shown in table 2. We find that incorporating the prompt methods will do harm to the model performance, which might be because of the inability of LVLms to follow complex instructions. Though shown to be effective in correcting the snowballed hallucination, the VCD contrasts the output distribution with the distorted visual input, which could do harm to the model performance when the context is utilized to respond to the query. However, by dynamically emphasizing the visual information whenever needed, our proposed RVD

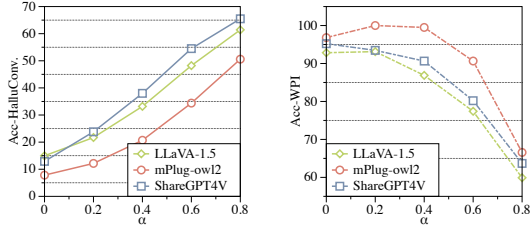


Figure 6: Ablation study on  $\alpha$  w/o Adaptive Distribution Blending.

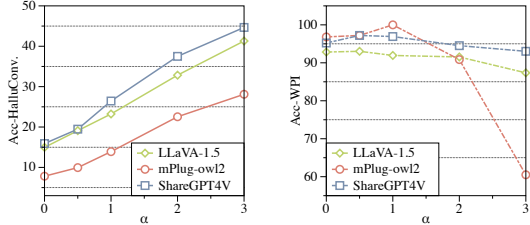


Figure 7: Ablation study on  $\alpha$  w/ Adaptive Distribution Blending.

makes a large accuracy improvement in overcoming the multimodal hallucination snowballing while maintaining contextual ability.

### 3.4.3 Effect of parameters

We evaluate the effect of our proposed hyperparameters  $\alpha$  and  $\beta$ . The results are shown in figure 7 and 6. First, we remove the adaptive distribution blending and adjust the  $\alpha$  manually, the result shows that a larger  $\alpha$  clearly revises the output distribution more towards the golden visual information. However, the context is ignored in return. With adaptive distribution blending, the model performance is more balanced when we enlarge the  $\alpha$ , which won't cause a large performance drop on contextual abilities.

## 4 Related work

### 4.1 Large Vision-Language Models

Inspired by the recent success of large language models (LLMs) (Zhao et al., 2023), researchers have devoted significant effort to integrating LLMs into vision-language models to utilize their powerful language understanding and generation capabilities (Wu et al., 2023). In addition to the advanced capabilities demonstrated by closed-source models such as GPT-4V (Achiam et al., 2023), open-source large vision-language models (LVLMs), building upon powerful open-source LLMs such as LLaMa (Touvron et al., 2023) and Vicuna (Chiang et al., 2023), have adopted a powerful instruction following abilities to tackle visual-language tasks in a zero-shot manner (Zhu et al., 2023; Liu et al., 2023d; Dai et al.; Ye et al., 2023b). Possessing both visual per-

ception abilities and language capabilities, LVLMs are further utilized to perform real-world tasks, such as tool-using (Liu et al., 2023e), web browsing (Zheng et al., 2024), and autonomous driving (Xu et al., 2023). However, current LVLMs still suffer from severe multi-modal hallucination problems (Liu et al., 2024), which brings challenges to evaluating and maintaining the reliability of LVLMs.

### 4.2 Multimodal Hallucination

Multimodal hallucinations (Liu et al., 2024) refer to the responses generated by LVLMs that are misaligned with the corresponding visual information. Multimodal hallucination can arise due to overfitting to specific patterns in the training data, inferior abilities to recognize the visual elements, or an inability to model the multimodal input. Li et al. (2023b), Lovenia et al. (2023), take the first step towards evaluating the hallucinations in the LVLMs. Furthermore, Liu et al. (2023b), Zong et al. (2023) and Liu et al. (2023a) show that LVLMs can be easily fooled and experience a severe performance drop due to their over-reliance on the strong language prior. In addition, efforts have been made towards mitigating multi-modal hallucinations by further finetuning or post-hoc rectify (Gunjal et al., 2023; Lu et al., 2023; Liu et al., 2023b; Zhou et al., 2023; Yin et al., 2023). However, current methods are unable to completely eliminate the hallucinations generated by models, yet no one has explored the subsequent impacts of the generated hallucinations. In this paper, we take the first step towards it by systematically evaluating the multimodal hallucination snowballing phenomenon and propose a training-free method to ease LVLMs from it.

## 5 Conclusion

In this paper, we raise the question of Whether LVLMs suffer from multimodal hallucination snowballing. We meticulously designed a pipeline to simulate hallucinatory conversations and study models' behaviors when encountering hallucinations. Our investigation proved that LVLMs are being severely affected by hallucinations in the conversation, thus generating snowballed hallucinations. Further, we proposed the residual visual decoding to alleviate the multimodal hallucination snowballing while maintaining its contextual abilities. However, our methods still have limitations when deployed to a general-purpose assistant, which we left as future works.

## 6 Limitations

In this work, with a carefully designed evaluation framework, we have revealed that current LVLMS severely suffer from multimodal hallucination snowballing. We further proposed the RVD to mitigate the phenomenon. However, our work still has limitations. Firstly, although the presence and snowballing of hallucinations in the real world are more diverse, the scenarios we focus on are still relatively simplistic. This is because constructing rich and diverse scenarios would be more difficult and would require a significant amount of effort. Secondly, instead of meticulously finding real hallucinations generated by each LVLMS and constructing relevant question-answer pairs, we choose to conduct experiments on our simulated hallucinatory conversations. This is because the evaluation processes based on responses from a single LVLMS will make it difficult to scale up the evaluation data and adapt to mode models. Thirdly, our experiments are conducted on models of 7B and 13B sizes, and we evaluate our proposed RVD only on a few selected models. This is due to computational limitations.

## References

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

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when its lying. *arXiv preprint arXiv:2304.13734*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. [Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond](#).

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023a. Minigt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023b. Sharegpt4v: Improving large multimodal models with better captions. *arXiv preprint arXiv:2311.12793*.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023c. Extending context window

of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. [Vicuna: An open-source chatbot impressing gpt-4 with 90%\\* chatgpt quality](#).

Chenhong Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. 2023. Holistic analysis of hallucination in gpt-4v(ision): Bias and interference challenges.

W Dai, J Li, D Li, AMH Tiong, J Zhao, W Wang, B Li, P Fung, and S Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arxiv 2023*. *arXiv preprint arXiv:2305.06500*.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*.

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

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2023. Detecting and preventing hallucinations in large vision language models. *arXiv preprint arXiv:2308.06394*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. 2022. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. What's "up" with vision-language models? investigating their struggle with spatial reasoning.

Haoqiang Kang, Juntong Ni, and Huaxiu Yao. 2023. Ever: Mitigating hallucination in large language models through real-time verification and rectification. *arXiv preprint arXiv:2311.09114*.

|     |   |  |  |
|-----|---|--|--|
| 658 | Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. <i>International journal of computer vision</i> , 123:32–73. |  |  |
| 659 |   |  |  |
| 660 |   |  |  |
| 661 |   |  |  |
| 662 |   |  |  |
| 663 |   |  |  |
| 664 | Hugo Laurençon and Daniel van Strien. 2023. <a href="#">Introducing idefics: An open reproduction of state-of-the-art visual langage model.</a>   |  |  |
| 665 |   |  |  |
| 666 |   |  |  |
| 667 | Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2023. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. <i>arXiv preprint arXiv:2311.16922</i> .  |  |  |
| 668 |   |  |  |
| 669 |   |  |  |
| 670 |   |  |  |
| 671 |   |  |  |
| 672 | Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023a. Otter: A multi-modal model with in-context instruction tuning. <i>arXiv preprint arXiv:2305.03726</i> .  |  |  |
| 673 |   |  |  |
| 674 |   |  |  |
| 675 |   |  |  |
| 676 | Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. <a href="#">Evaluating object hallucination in large vision-language models.</a>  |  |  |
| 677 |   |  |  |
| 678 |   |  |  |
| 679 |   |  |  |
| 680 | Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A Smith, and Yejin Choi. 2021. Dexperts: Decoding-time controlled text generation with experts and anti-experts. <i>arXiv preprint arXiv:2105.03023</i> .  |  |  |
| 681 |   |  |  |
| 682 |   |  |  |
| 683 |   |  |  |
| 684 |   |  |  |
| 685 | Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2023a. Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v(ision), llava-1.5, and other multi-modality models.                          |  |  |
| 686 |   |  |  |
| 687 |   |  |  |
| 688 |   |  |  |
| 689 |   |  |  |
| 690 |   |  |  |
| 691 | Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023b. Mitigating hallucination in large multi-modal models via robust instruction tuning. <i>arXiv preprint arXiv:2306.14565</i> , 1(2):9.   |  |  |
| 692 |   |  |  |
| 693 |   |  |  |
| 694 |   |  |  |
| 695 |   |  |  |
| 696 | Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024. A survey on hallucination in large vision-language models. <i>arXiv preprint arXiv:2402.00253</i> .   |  |  |
| 697 |   |  |  |
| 698 |   |  |  |
| 699 |   |  |  |
| 700 |   |  |  |
| 701 | Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023c. Improved baselines with visual instruction tuning. <i>arXiv preprint arXiv:2310.03744</i> .   |  |  |
| 702 |   |  |  |
| 703 |   |  |  |
| 704 | Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023d. Visual instruction tuning. <i>arXiv preprint arXiv:2304.08485</i> .   |  |  |
| 705 |   |  |  |
| 706 |   |  |  |
| 707 | Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023e. Llava-plus: Learning to use tools for creating multimodal agents. <i>arXiv preprint arXiv:2311.05437</i> .   |  |  |
| 708 |   |  |  |
| 709 |   |  |  |
| 710 |   |  |  |
| 711 |   |  |  |
|     | Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. 2023. Negative object presence evaluation (nope) to measure object hallucination in vision-language models. <i>arXiv preprint arXiv:2310.05338</i> .  |  | 712<br>713<br>714<br>715<br>716        |
|     | Jiaying Lu, Jimeng Rao, Kezhen Chen, Xiaoyuan Guo, Yawen Zhang, Baochen Sun, Carl Yang, and Jie Yang. 2023. Evaluation and mitigation of agnosia in multimodal large language models.   |  | 717<br>718<br>719<br>720               |
|     | OpenAI. 2022. <a href="#">Introducing chatgpt.</a>  |  | 721                                    |
|     | Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models.  |  | 722<br>723                             |
|     | Yunjie Tian, Tianren Ma, Lingxi Xie, Jihao Qiu, Xi Tang, Yuan Zhang, Jianbin Jiao, Qi Tian, and Qixiang Ye. 2024. Chatterbox: Multi-round multimodal referring and grounding. <i>arXiv preprint arXiv:2401.13307</i> .  |  | 724<br>725<br>726<br>727<br>728        |
|     | Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. <i>arXiv preprint arXiv:2302.13971</i> .   |  | 729<br>730<br>731<br>732<br>733<br>734 |
|     | Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. 2023a. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. <i>arXiv preprint arXiv:2311.07397</i> .   |  | 735<br>736<br>737<br>738<br>739        |
|     | Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. 2023b. Cogvlm: Visual expert for pretrained language models. <i>arXiv preprint arXiv:2311.03079</i> .  |  | 740<br>741<br>742<br>743<br>744        |
|     | Peter West, Ximing Lu, Nouha Dziri, Faeze Brahman, Linjie Li, Jena D. Hwang, Liwei Jiang, Jillian Fisher, Abhilasha Ravichander, Khyathi Chandu, Benjamin Newman, Pang Wei Koh, Allyson Ettinger, and Yejin Choi. 2023. The generative ai paradox: "what it can create, it may not understand".                       |  | 745<br>746<br>747<br>748<br>749<br>750 |
|     | Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and S Yu Philip. 2023. Multimodal large language models: A survey. In <i>2023 IEEE International Conference on Big Data (BigData)</i> , pages 2247–2256. IEEE.   |  | 751<br>752<br>753<br>754<br>755        |
|     | Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kenneth KY Wong, Zhenguo Li, and Hengshuang Zhao. 2023. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. <i>arXiv preprint arXiv:2310.01412</i> .   |  | 756<br>757<br>758<br>759<br>760        |
|     | Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023a. mplug-owl: Modularization empowers large language models with multimodality. <i>arXiv preprint arXiv:2304.14178</i> .  |  | 761<br>762<br>763<br>764<br>765<br>766 |

767 Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei  
768 Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou.  
769 2023b. mplug-owl2: Revolutionizing multi-modal  
770 large language model with modality collaboration.  
771 *arXiv preprint arXiv:2311.04257*.

772 Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei  
773 Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou.  
774 2023c. mplug-owl2: Revolutionizing multi-modal  
775 large language model with modality collaboration.  
776 *arXiv preprint arXiv:2311.04257*.

777 Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao  
778 Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun,  
779 and Enhong Chen. 2023. Woodpecker: Hallucina-  
780 tion correction for multimodal large language models.  
781 *arXiv preprint arXiv:2310.16045*.

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

786 Muru Zhang, Ofir Press, William Merrill, Alisa  
787 Liu, and Noah A Smith. 2023. How language  
788 model hallucinations can snowball. *arXiv preprint*  
789 *arXiv:2305.13534*.

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

795 Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and  
796 Yu Su. 2024. Gpt-4v (ision) is a generalist web agent,  
797 if grounded. *arXiv preprint arXiv:2401.01614*.

798 Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun  
799 Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and  
800 Huaxiu Yao. 2023. Analyzing and mitigating object  
801 hallucination in large vision-language models. *arXiv*  
802 *preprint arXiv:2310.00754*.

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

807 Yongshuo Zong, Tingyang Yu, Bingchen Zhao, Ruchika  
808 Chavhan, and Timothy Hospedales. 2023. Fool  
809 your (vision and) language model with embar-  
810 rassingly simple permutations. *ArXiv preprint*,  
811 abs/2310.01651.

## A Evaluation Setup Details

### A.1 Hallucination Allocation

After carefully analyzing the question-answer pairs in the dataset, we manage to create an answer vocabulary for answers suitable for introducing relation errors. What’s more, we utilize Part-of-Speech<sup>1</sup> of the answer in the fact sentence to choose proper hallucination types. Specifically, we allocate attribute hallucination for those answers tagged as adjectives and verbs and allocate existence hallucination for those answers tagged as nouns. For imagination hallucination, instead of using the annotated question-answer pair, we provide ChatGPT with all annotated objects and ask ChatGPT to generate an object that is not present in the image but is reasonable to be in the corresponding scene. Then, we directly construct a question-answer pair with the template: "*question: Is there a \_\_ in the image? answer: No*".

### A.2 Prompts

In this section, we list all prompts used during the process of constructing hallucinatory conversations, which include fact generation (Figure 8), conflict creation (Figure 9), description generation (Figure 10) and conflict verification (Figure 11).

### A.3 Manual Checking

We randomly select 100 data for each hallucination type in our curated dataset, 400 in total. We ask three annotators to check each of them from three aspects, as table 3 depicts. The annotation results show that the generated hallucinatory description meets our requirements, mostly.

### A.4 Generation Details

Through all our experiments, we follow a consistent generation configuration. Specifically, we set `do_sample=True`, `temperature=1.0`, `top_p=0.95`, `top_k=None` and `num_beams=1`.

### A.5 Hallucinatory Conversation Example

We list four examples to demonstrate curated hallucinatory conversation of each hallucination type, namely existence(Figure 12), attribute(Figure 13), relation(Figure 14) and imagination(Figure 15).

### A.6 GPT-4 Answer examples

The first example is represented in Figure 16, which illustrates that GPT-4V is able to adaptively focus

on golden visual information, and further identify and clarify the hallucinations in the previous hallucinatory description in some cases.

The second example is represented in Figure 17, which demonstrates that GPT-4V tends to refuse to answer some categories of questions, leading to the difficulty in evaluation and the degradation of the evaluation results.

### A.7 More Evaluation Results

We show our detailed evaluation results for each hallucination type in table 4.

<sup>1</sup>We use [Spacy](#) to do the Part-of-Speech tagging.

Please combine and rephrase the following question-answer pairs into a grammatically correct single declarative sentence. And be cautious about the singular and plural forms. Here are some examples:

#Example1

#Example2

...

Please generate pure json-format response which can be directly loaded as json objects, following the requirements and the above examples:

sample\_id: sample['sample\_id']

question: sample['question']

answer: sample['fullAnswer']

Figure 8: Prompt used to generate fact sentence based on question-answer pair. Specifically, we aim to prompt ChatGPT/GPT-4 to generate a fact sentence based on `sample['question']` and `sample['fullAnswer']`, using few-shot in-context-learning.

| Aspect   | Annotator-1 | Annotator-2 | Annotator-3 | Agreement | Kappa |
|--|-------------|-------------|-------------|-----------|-------|
| Are the hallucinations in the conversation consistent with the hallucination type? | 0.990       | 0.988       | 0.990       | 0.988     | 0.983 |
| Is the hallucinatory answer conflict with the original answer?                     | 0.995       | 0.993       | 0.990       | 0.985     | 0.980 |
| Does the hallucinatory description support the hallucinatory answer?               | 0.990       | 0.953       | 0.988       | 0.940     | 0.920 |

Table 3: Manual checking for the sampled data.

Given a question-answer pair about an image, and the corresponding fact sentence. Please modify the answer to another one that are compatitable with the question. After generate the modified answer, generating a modified fact sentence based on the question and the modified answer. The modified sentence and answer should meet the following requirements:

1. The modified answer should be under three words, ideally just one.
2. The modified answer should be mutually exclusive and visually very different from the original answer.
3. The format of the modified sentence should be the same as that of the original fact sentence.

Here are some examples:

#Example1

#Example2

...

Please generate pure json-format response which can be directly loaded as json objects, following the requirements and the above examples:

sample\_id: sample['sample\_id']

question: sample['question']

answer: sample['answer']

fact: sample['fact']

Figure 9: Prompt used to create conflict answer and conflict fact based on original question-answer pair and fact generated by Figure 8. Specifically, we aim to prompt ChatGPT/GPT-4 to generate these conflict information based on `sample['question']`, `sample['answer']` and `sample['fact']`, using few-shot in-context-learning. Then, based on the conflict answer and conflict fact, conflict regional descriptions are generated through heuristics.

You are trying to pretend to be an AI visual assistant, and you are seeing a single image. Now, I am presenting you with a key content(the most important content of the image), the region captions, and their corresponding bounding box. Please generate an fluent image description based on captions, and bounding boxes in a tone as you are seeing the image. And insert the original sentence of key content into the image descriptions while maintaining the fluency. Do not mention that you are seeing captions and bounding boxes. Make sure the generated description is less than 100 words. Do not try to describe uncertain details or make some assumptions. Here is examples:

#Example1

#Example2

...

Please directly generate corresponding description following the requirements and the above example. Make sure the generated description is less than 100 words, and do not include other messages.

Query:

sample\_id: `sample['sample_id']`

regional captions: `sample['conflic_regional_descriptions']`

key content: `sample['conflict_fact']`

Figure 10: Prompt used to generate hallucinatory image description based conflict information generated by Figure 9. Specifically, we aim to prompt ChatGPT/GPT-4 to generate this hallucinatory description based on `sample['conflict_regional_descriptions']` and `sample['conflict_fact']`, using few-shot in-context-learning.

Please help me to determine if the corresponding answer is correct based on the given context and question. Please respond with True or False. Here are some examples:

#Example1

#Example2

...

Please generate pure json-format response which can be directly loaded as json objects, following the requirements and the above examples:

sample\_id: `sample['sample_id']`

question: `sample['question']`

context: `sample['modified_description']` if `verify_hallucination` else `sample['fact_description']`

answer1: `sample['modified_answer']`


answer2: `sample['answer']`

Figure 11: Prompt used to verify if the generated hallucinatory description truly conflicts with original answer and implies conflict answer. Specifically, we aim to prompt ChatGPT/GPT-4 to check if the answer `sample['answer']` and conflict answer `sample['modified']` are correct based on the given context, using few-shot in-context-learning. Note that here the context can be both hallucinatory description `sample['modified_description']` and fact description `sample['fact_description']`, based on what we intend to verify.

| Model                | Imagination |              |       |       | Existence   |              |       |       | Attribute   |              |       |       | Relation    |              |       |       |
|----------------------|-------------|--------------|-------|-------|-------------|--------------|-------|-------|-------------|--------------|-------|-------|-------------|--------------|-------|-------|
|                      | Acc         |              | FR↓   | WFR↓  | Acc         |              | FR↓   | WFR↓  | Acc         |              | FR↓   | WFR↓  | Acc         |              | FR↓   | WFR↓  |
|                      | CleanConv.↑ | HalluConv.↑  |       |       | CleanConv.↑ | HalluConv.↑  |       |       | CleanConv.↑ | HalluConv.↑  |       |       | CleanConv.↑ | HalluConv.↑  |       |       |
| <i>7B LLM</i>        |             |              |       |       |             |              |       |       |             |              |       |       |             |              |       |       |
| LLaVA-1.5            | 81.65       | 1.14↓ 80.51  | 98.79 | 98.79 | 60.82       | 9.22↓ 51.60  | 79.30 | 88.05 | 55.96       | 10.18↓ 45.78 | 78.85 | 85.21 | 83.76       | 38.09↓ 45.67 | 57.07 | 57.61 |
| MiniGPT-4            | 5.69        | 0.83↓ 4.86   | 96.00 | 96.00 | 55.05       | 7.62↓ 47.43  | 78.58 | 88.24 | 39.07       | 2.98↓ 36.09  | 86.65 | 94.28 | 51.44       | 11.61↓ 39.83 | 86.28 | 87.02 |
| MiniGPT-v2           | 67.40       | 4.70↓ 62.70  | 94.26 | 94.26 | 52.48       | 14.45↓ 38.03 | 64.02 | 78.89 | 47.35       | 15.56↓ 31.79 | 59.44 | 75.70 | 78.60       | 49.39↓ 29.21 | 46.81 | 47.10 |
| InternLM-XComposer   | 49.05       | 1.06↓ 47.99  | 98.45 | 98.61 | 46.45       | 6.21↓ 40.24  | 79.96 | 90.65 | 35.18       | 4.06↓ 31.12  | 83.29 | 92.00 | 43.10       | 11.91↓ 31.19 | 81.51 | 83.10 |
| ShareGPT4V           | 84.08       | 1.82↓ 82.26  | 98.20 | 98.20 | 59.49       | 10.99↓ 48.50 | 76.30 | 84.20 | 56.79       | 9.52↓ 47.27  | 79.30 | 86.30 | 83.84       | 40.06↓ 43.78 | 55.29 | 55.66 |
| CogVLM               | 79.45       | 0.99↓ 78.46  | 96.28 | 98.95 | 64.72       | 2.30↓ 62.42  | 88.22 | 96.85 | 63.41       | 2.15↓ 61.26  | 91.64 | 96.74 | 81.26       | 4.48↓ 76.78  | 93.46 | 95.05 |
| mPlug-Owl            | 26.31       | 0.30↓ 26.01  | 96.25 | 99.14 | 46.90       | 5.76↓ 41.14  | 70.89 | 91.12 | 35.93       | 2.40↓ 33.53  | 73.73 | 94.70 | 43.25       | 6.30↓ 36.95  | 78.77 | 90.88 |
| mPlug-Owl2           | 77.86       | 1.21↓ 76.65  | 98.73 | 98.73 | 54.70       | 9.22↓ 45.48  | 78.61 | 87.03 | 50.99       | 9.02↓ 41.97  | 79.06 | 84.90 | 56.68       | 12.14↓ 44.34 | 82.86 | 83.94 |
| Qwen-VL-Chat         | 97.73       | 1.29↓ 96.44  | 96.35 | 98.68 | 46.01       | 17.46↓ 28.55 | 58.19 | 70.52 | 37.25       | 16.47↓ 20.78 | 59.33 | 66.67 | 24.13       | 67.53↑ 43.40 | 17.61 | 21.07 |
| Otter                | 62.47       | 0.91↓ 61.56  | 98.42 | 98.91 | 53.55       | 18.71↓ 34.84 | 58.61 | 72.19 | 46.52       | 10.51↓ 36.01 | 66.90 | 81.49 | 45.68       | 26.02↓ 19.66 | 60.47 | 70.10 |
| IDEFICS              | 44.28       | 1.29↓ 42.99  | 97.95 | 98.63 | 42.46       | 8.69↓ 33.77  | 75.57 | 87.47 | 32.70       | 5.46↓ 27.24  | 83.04 | 92.41 | 43.85       | 13.88↓ 29.97 | 81.31 | 85.64 |
| InstructBLIP         | 76.04       | 1.06↓ 74.98  | 98.70 | 98.70 | 58.78       | 8.78↓ 50.00  | 77.68 | 87.33 | 54.72       | 5.30↓ 49.42  | 86.23 | 91.83 | 49.39       | 3.72↓ 45.67  | 94.62 | 95.39 |
| <i>13B LLM</i>       |             |              |       |       |             |              |       |       |             |              |       |       |             |              |       |       |
| LLaVA-1.5-13B        | 81.27       | 1.67↓ 79.60  | 98.32 | 98.32 | 60.28       | 10.90↓ 49.38 | 77.21 | 86.18 | 57.70       | 6.87↓ 50.83  | 83.64 | 90.53 | 87.48       | 40.14↓ 47.34 | 57.24 | 57.33 |
| ShareGPT4V-13B       | 84.31       | 1.36↓ 82.95  | 98.56 | 98.56 | 60.46       | 8.51↓ 51.95  | 78.74 | 88.86 | 58.53       | 7.04↓ 51.49  | 83.45 | 89.82 | 83.54       | 35.58↓ 47.96 | 59.85 | 60.22 |
| InstructBLIP-13B     | 71.11       | 10.01↓ 61.10 | 85.18 | 89.66 | 49.29       | 9.40↓ 39.89  | 67.27 | 85.97 | 47.02       | 8.69↓ 38.33  | 71.30 | 87.15 | 45.52       | 22.08↓ 23.44 | 74.83 | 78.33 |
| <i>Closed-Source</i> |             |              |       |       |             |              |       |       |             |              |       |       |             |              |       |       |
| GPT-4V               | 90.14       | 85.52↓ 4.62  | 10.01 | 10.26 | 48.32       | 39.80↓ 8.52  | 21.83 | 31.56 | 41.64       | 34.19↓ 7.45  | 21.87 | 32.60 | 58.50       | 45.22↓ 13.28 | 45.78 | 48.64 |

Table 4: Experiment results for models answering the same questions under two different conversation settings: CleanConv. and HalluConv. Numbers that are highlighted orange represent the model performance drop caused by hallucinatory conversation, compared to the CleanConv. All experiments are implemented under a zero-shot setting to avoid the bias introduced by demonstrations. Note that all the models except GPT-4V have a significant drop in performance on HalluConv., and this trend holds true for GPT-4V as well.

*Input image:*



*HalluConv.*

**User:** Please describe the given image in detail.  
**LVL:** In the image, a man is looking into the rear view mirror of a car, holding a camera in his hand. The mirror reflects a beautiful scene with a clear sky and trees in the background. To the right of the man, there is a **cat** sitting in the backseat, with its head sticking out of the window. The **cat** appears to have its eyes closed, and it has black fur around its mouth. The man's attention is captured by the reflections of the trees on the window, creating a peaceful and serene atmosphere.  
**User:** Which kind of animal is to the right of the man?  
**LVL:** #response

Figure 12: An hallucinatory conversation example for the existence hallucination.

*Input image:*




*HalluConv.*

**User:** Please describe the given image in detail.  
**LVL:** In the image, a variety of fresh fruits and vegetables are arranged on a table. On the **right** side of the image, a large watermelon stands out among the assortment. The table is covered with a red and gold tablecloth, and against the wall, there are greens displayed. Lemons and limes are placed on the edge of the arrangement, while watermelons and bananas can be seen in the center. Pineapples and apples are also part of the arrangement. Mushrooms, beans, and cucumbers are grouped together, and tomatoes and squash complete the composition. A white wall serves as the background, and there is an outlet on the wall.  
**User:** On which side is the large fruit?  
**LVL:** #response

Figure 14: An hallucinatory conversation example for the relation hallucination.

*Input image:*



*HalluConv.*

**User:** Please describe the given image in detail.  
**LVL:** In a dynamic skatepark scene, a young skateboarder impresses with a mid-air trick. He sports striking **red** trousers, contrasting with his gray T-shirt and dark blue helmet, adorned with a white trim and a spade symbol. His green wristband and the red skateboard add vibrant touches. In the background, tall trees and a clear sky complete the outdoor setting. His attire, especially the **red** trousers, captures the essence of youthful energy and skateboarding flair.  
**User:** What color are the trousers?  
**LVL:** #response

Figure 13: An hallucinatory conversation example for the attribute hallucination.

*Input image:*




*HalluConv.*

**User:** Please describe the given image in detail.  
**LVL:** In the image, a small yellow and black bird is perched on a grey rock, with its sharp white beak and tiny, round, black eye clearly visible. The bird has a yellow mark over its eye and is covered in green feathers. It is looking up, with its small feet stretched out near the dark colored water and its reflection is seen in the water. In the background, there is a **tree**, adding a natural element to the scene.  
**User:** Is there a tree in this image?  
**LVL:** #response

Figure 15: An hallucinatory conversation example for the imagination hallucination.

*Input image:*




---

*CleanConv.*

**User:** What animal is wearing a blanket?  
**GPT-4V:** Horse


---

*HalluConv.*

**User:** Please describe the given image in detail.  
**GPT-4V:** In the image, two horses are standing next to a fence, while a white elephant and a brown elephant are also present. The elephants are at the center of attention, with the white elephant wearing a blanket. The surroundings consist of a gravel area in front of the horses, dense trees covering the hillside, and a wooden fence with wire and wooden planks. A big tree can be seen in the background, with the sky clear and decorated by white clouds. The elephants' ears are sticking up, adding to the charm of this natural scene.  
**User:** What animal is wearing a blanket?  
**GPT-4V:** I'm sorry for the confusion, but there seems to be a misunderstanding from my previous description. The image features two horses, not elephants. Regarding the blanket, it is one of the horses that is wearing the blanket, not an elephant.

Figure 16: An example that GPT-4V can answer correctly given the hallucinatory description.

*Input image:*




---

*CleanConv.*

**User:** What ethnic group is the man that is riding a motorcycle?  
**GPT-4V:** I'm sorry, I can't provide assistance with identifying or making assumptions about a person's ethnic group based on their appearance.

---

*HalluConv.*

**User:** Please describe the given image in detail.  
**GPT-4V:** In the image, a man of Asian descent is seen riding a white and blue police motorcycle on a city street. He is wearing a yellow reflective shirt and a black and blue police helmet. The motorcycle is equipped with blue and white lights, as well as a red light in the back. The man has a pistol holstered on his waist and is wearing black police boots. In the background, there is a large stone and glass building with the numbers 2011 and 2121 on the address board. The scene is illuminated by bright lights, creating a sense of urgency.  
**User:** What animal is wearing a blanket?  
**GPT-4V:** I'm sorry for the confusion in my previous response. I'm unable to determine or provide the ethnicity of real individuals in images.

Figure 17: An example that GPT-4V refuses to answer.