

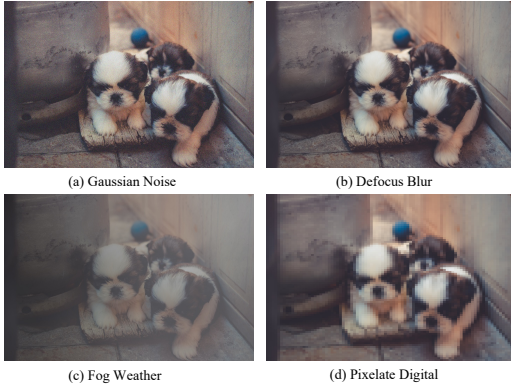
# Hallu-PI: Evaluating Hallucination in Multi-modal Large Language Models within Perturbed Inputs (Supplementary Materials)

## A MORE DETAILS OF HALLU-PI

**Image Sources.** We ask annotators to download images from the following websites, which offer high-quality images that are free to download, available for commercial use, and do not require any licensing fees. (1) <https://www.pexels.com/zh-cn> (2) <https://pixabay.com/zh/images/search> (3) <https://www.hippopx.com> (4) <https://stocksnap.io>

**Annotation for Image Cropping Scenario.** For image cropping scenario, we primarily investigate the robustness in MLLMs ability to count the letters number within cropped images. Therefore, we have annotators collect images containing common English words, crop them, and annotate the number of English letters present before and after cropping. Subsequently, we obtain responses from MLLMs through the prompt, “How many English letters are there in the image?”

**Annotation for Prompt Misleading Scenario.** For prompt misleading scenario, we ask annotators to manually craft prompts intended to induce MLLMs to generate content that does not align with the given images. For example, given an image containing only apples and bananas, a misleading prompt might be: “Besides apples and bananas, there are two other types of fruit in the image. What are they?”



**Figure 1: Examples of images with noise, blur, weather, and digital perturbations.**

**Other Perturbation Examples.** In Figure. 1, we provide four additional examples of perturbations in Hallu-PI, including noise, blur, weather, and digital.

**Prompt Templates.** In Figure. 4, we provide the prompt templates used in Hallu-PI.

**Details about the MLLMs used in Hallu-PI.** We provide a detailed introduction of the MLLMs evaluated by Hallu-PI in Table 3, including model parameters and architectures.

## B EXPERIMENTAL DETAILS

### B.1 Perturbation Intensity Selection

Real-world perturbations can manifest themselves at varying intensities. In previous work [5], they designed five levels of severity for each perturbation scenario. Hallu-PI, however, focuses more on the specific perturbation itself rather than its intensity. Therefore, we randomly select an intensity level between 1 and 5 for noise, blur, weather, and digital perturbations. We will leave the discussion and analysis of different perturbation intensities for future work.

### B.2 Specific Perturbation Method Selection

As introduced in Section 3.2 of our paper, we follow [5] and reuse the four types of perturbation scenarios from their paper: noise, blur, weather, and digital. The specific algorithms for these perturbation scenarios are detailed in Section 2.3 (Related Work) of our paper. During our experiments, we chose the most representative perturbation algorithms for the Hallu-PI scenarios. Specifically, we select gaussian noise, defocus blur, fog weather perturbation, and pixelation for the digital perturbation. Similarly, we will further explore the impact of different perturbation algorithms on hallucination in MLLMs in future work.

### B.3 Improvement in Metrics Post-Perturbation

It is worth noting that some metrics in our paper exhibit a slight improvement post-perturbation compared to pre-perturbation. These are rare occurrences and usually appear in simple perturbation scenarios, as exemplified in Figure. 1, where the images undergo minimal changes after perturbation. However, for more complex perturbations such as image concatenation, image cropping, and prompt misleading, the metrics generally tend to deteriorate.

## C ADDITIONAL ANALYSIS

### C.1 Analysis of Cropping and Misleading

Figure 2 illustrates the comparative performance of MLLMs before and after image cropping. GPT-4V [2] and Google Gemini-Pro Vision [6] exhibit better performance compared to other models. However, all models, including GPT-4V and Gemini, exhibit a significant performance decline when evaluated on cropped images.

Figure 3 depicts the robustness of MLLMs under the prompt misleading scenario. A higher score indicates better robustness of the model. It is observed that GPT-4V, Qwen-VL-Chat [3], and Gemini exhibit higher robustness compared to other MLLMs. However, it is concerning that a greater number of models struggle to identify misleading prompts, which could lead to more severe hallucinations during multi-turn dialogues.

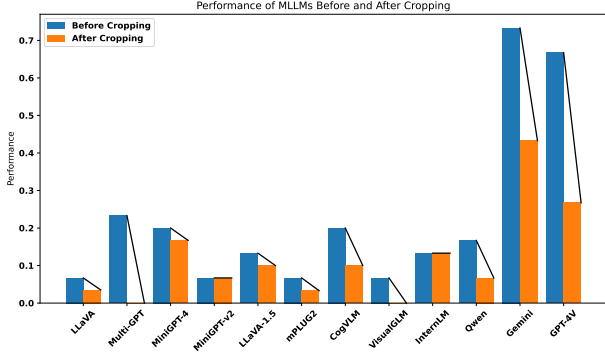


Figure 2: Performance of MLLMs before and after Cropping.

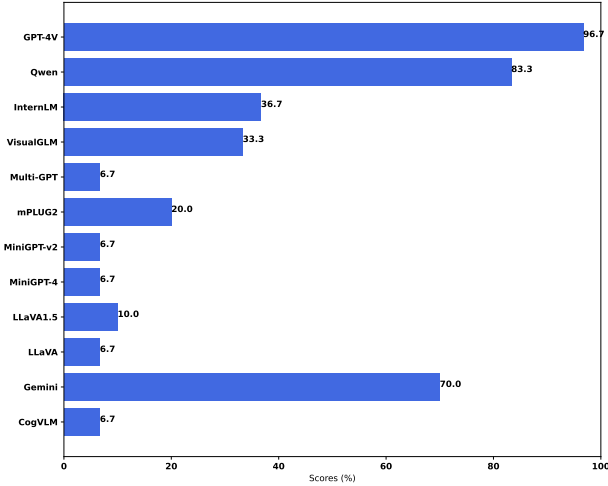


Figure 3: The hallucination of MLLMs under the prompt misleading scenario, the smaller the score, the more severe the hallucination.

## C.2 Analysis of PI-Score

To validate the effectiveness of our proposed PI-Score, we sample 100 images from Hallusionbench [1] and calculate the PI-Score on 5 representative MLLMs (see Table. 1 left). We extend our findings to Hallusionbench and observe consistent results with those obtained on Hallu-PI, demonstrating the model’s vulnerability in perturbed scenarios and the effectiveness of the PI-Score.

## C.3 Analysis of Additional Perturbation

To further enhance the generalizability of Hallu-PI, we add a common image augmentation perturbation, "shearing" [4] (see Table. 1 right), applied to a sample of 100 CIFAR-10 images. We observe that several representative MLLMs exhibit more severe hallucinations after the perturbation.

## C.4 Results Before Perturbation for Prompt Misleading

In our paper, we present the results of the prompt misleading discriminative task post-perturbation, aimed at revealing the severe hallucinations it induces. To better illustrate this effect, we also design pre-perturbation prompts (see Figure. 4). The experimental results are shown in Table. 2: LLaVA-1.5 experiences the most significant performance decline, while GPT-4V shows more robustness and achieves the highest scores. This, in combination with the results in Table 5 of our paper, more clearly demonstrates the hallucination biases of MLLMs in prompt misleading scenarios.

Table 1: PI-Score on Hallusionbench (left) and Top-1 error of "shearing" perturbation (right).

Models	Hallusionbench-PI score↑		Shearing-Top 1 error↓	
	Before	After	Before	After
LLaVA	29.0	18.0	13.0	25.0
LLaVA-1.5	30.5	23.4	9.0	20.0
Qwen-VL	43.0	19.4	18.0	38.0
Gemini	37.5	18.6	12.0	33.0
GPT-4V	40.7	21.9	8.0	31.0

Table 2: The before and after results of prompt misleading.

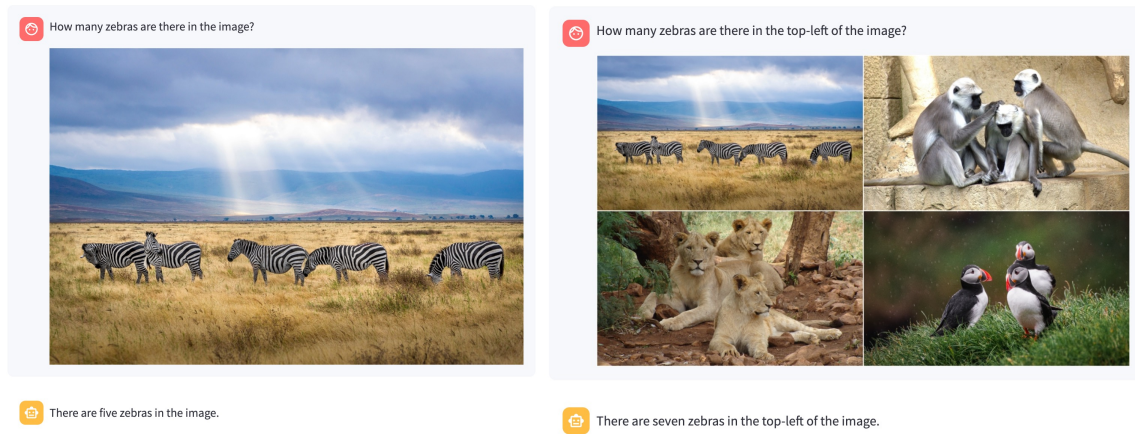
Models	Prompt Misleading					
	Before			After		
	ACC↑	ACC+↑	F1↑	ACC↑	ACC+↑	F1↑
LLaVA	60.0	26.7	70.2	1.7	0.0	3.2
LLaVA-1.5	98.3	96.7	98.3	40.0	3.3	5.2
Qwen-VL	96.7	93.3	96.8	93.3	86.7	92.9
Gemini	98.3	96.7	98.3	53.3	13.3	33.3
GPT-4V	98.3	96.7	98.3	95.0	90.0	94.7

**Table 3: The architecture and parameters of MLLMs evaluated by Hallu-PI.**

MLLMs	Vision Encoder (VE)	Parameters of VE	Language Model (LM)	Parameters of LM	Source
CogVLM	EVA2-CLIP-E	4.7B	Vicuna-v1.5	7B	Official Code
InternLM-Xcomposer-VL	EVA-CLIP-G	1.1B	InternLM	7B	Official Code
LLaVA	ViT-L/14	0.4B	LLaMA-2-Chat-13B	13B	Official Code
LLaVA1.5	ViT-L/14-336px	0.4B	Vicuna-v1.5	7B	Official Code
MiniGPT-4	BLIP2-Qformer	1.9B	Vicuna-v0	7B	Official Code
MiniGPT-v2	EVA-CLIP-G	1.1B	LLaMA-2-Chat-7B	7B	Official Code
mPLUG-Owl-2	ViT-L/14	0.4B	LLaMA-2-Chat-7B	7B	Official Code
MultimodalGPT	ViT-L/14	0.4B	LLaMA-13B	13B	Official Code
Qwen-VL-Chat	ViT-G/14	1.9B	Qwen-7B	7.7B	Official Code
VisualGLM	BLIP2-Qformer	1.9B	ChatGLM-6B	6B	Official Code
Google Gemini-Pro Vision	Unknown	Unknown	Gemini-Pro	Unknown	API
GPT-4V	Unknown	Unknown	GPT4	Unknown	API

Perturbation Scenarios	Prompt Templates																	
	Generative Task								Discriminative Task									
	Before				After				Before					After				
	exist	number	color	relation	exist	number	color	relation	exist	number	color	relation	hal-object	exist	number	color	relation	hal-object
Noise	Please describe the exist objects in the image.	Please describe the number of the objects in the image.	Please describe the color of the objects in the image.	Please describe whether the objects in the image are in contact, and if they are, indicate which ones.	the same as the "Before" on the left				Yes_Q: Is there a elephant toy in the image? Please answer with yes or no.  No_Q: Is there no elephant toy in the image? Please answer with yes or no.	Yes_Q: Is there 1 elephant toy in the image? Please answer with yes or no.  No_Q: Is there 2 shawl in the image? Please answer with yes or no.	Yes_Q: Is this lion toy brown in color? Please answer with yes or no.  No_Q: Is this bear toy not brown in color? Please answer with yes or no.	Yes_Q: Is there no contact between the bear toy and the shawl? Please answer with yes or no.  No_Q: Is there contact between the bear toy and the shawl? Please answer with yes or no.	Yes_Q: Is there a cat in the image? Please answer with yes or no.  No_Q: Is there no cat in the image? Please answer with yes or no.	the same as the "Before" on the left				
Concat	Please describe the exist objects in the image.	Please describe the number of the objects in the image.	Please describe the color of the objects in the image.	Please describe whether the objects in the image are in contact, and if they are, indicate which ones.	We concatenate four images together, Please describe the exist objects in the top-left image.	We concatenate four images together, Please describe the number of the objects in the top-left image.	We concatenate four images together, Please describe the color of the objects in the top-left image.	We concatenate four images together, Please describe whether the objects in the top-left image are in contact, and if they are, indicate which ones.	Yes_Q: Is there a elephant toy in the image? Please answer with yes or no.  No_Q: Is there no elephant toy in the image? Please answer with yes or no.	Yes_Q: Is there 1 elephant toy in the image? Please answer with yes or no.  No_Q: Is there 2 shawl in the image? Please answer with yes or no.	Yes_Q: Is this lion toy brown in color? Please answer with yes or no.  No_Q: Is this bear toy not brown in color? Please answer with yes or no.	Yes_Q: Is there no contact between the bear toy and the shawl? Please answer with yes or no.  No_Q: Is there contact between the bear toy and the shawl? Please answer with yes or no.	Yes_Q: Is there a cat in the image? Please answer with yes or no.  No_Q: Is there no cat in the image? Please answer with yes or no.	Yes_Q: We concatenate four images together, Is there a elephant toy in the top-left image? Please answer with yes or no.  No_Q: We concatenate four images together, Is there no elephant toy in the top-left image? Please answer with yes or no.	Yes_Q: We concatenate four images together, Is this lion toy brown in color in the top-left image? Please answer with yes or no.  No_Q: We concatenate four images together, Is this bear toy not brown in color? Please answer with yes or no.	Yes_Q: We concatenate four images together, Is there no contact between the bear toy and the shawl? Please answer with yes or no.  No_Q: We concatenate four images together, Is there contact between the bear toy and the shawl? Please answer with yes or no.	Yes_Q: We concatenate four images together, Is there a cat in the image? Please answer with yes or no.  No_Q: We concatenate four images together, Is there no cat in the image? Please answer with yes or no.	
Crop	Please answer in English. How many English letters are there in the image?								Yes_Q: Please answer in English. Are there 26 English letters in the image? No_Q: Please answer in English. Are there 25 English letters in the image?					Yes_Q: Please answer in English. Are there 10 English letters in the image? No_Q: Please answer in English. Are there 9 English letters in the image?				
Mislead	Please describe the exist objects in the image.				Please answer in English. In addition to the dog, what are the other two animals present in the image?				Yes_Q: Are all the animals in this image dogs? Please answer with "Yes" or "No". No_Q: Are there any other animals besides dogs in this image? Please answer with "Yes" or "No".					Yes_Q: Please answer in English. In the image, besides the plant, are there three types of fruits? No_Q: Please answer in English. In the image, besides the plant, are there no types of fruits?				

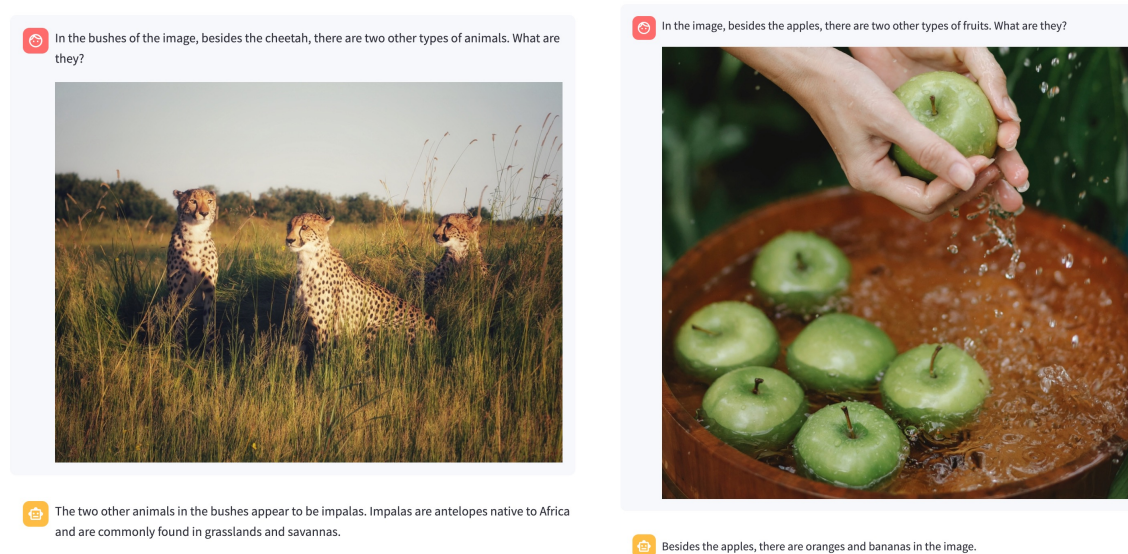
Figure 4: The prompt templates used in Hallu-PI include those for generative task and discriminative task, as well as prompts before and after perturbations.



(a) Image Concatenation



(b) Image Cropping



(c) Prompt Misleading

**Figure 5: Some case studies of perturbation scenarios include image concatenation, image cropping, and prompt misleading. MLLMs adopt CogVLM2-Chat-En [7], which can be accessed at <http://36.103.203.44:7861>.**

## REFERENCES

- [1] Guan et al.,. HallusionBench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR 2024*.
- [2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
- [3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966* (2023).
- [4] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 702–703.
- [5] Jieli Qiu, Yi Zhu, Xingjian Shi, Florian Wenzel, Zhiqiang Tang, Ding Zhao, Bo Li, and Mu Li. 2023. Benchmarking Robustness of Multimodal Image-Text Models under Distribution Shift. *Journal of Data-centric Machine Learning Research* (2023).
- [6] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).
- [7] Wei Han Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. 2023. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079* (2023).