Supplemental Material for Who Brings the Frisbee: Probing Hidden Hallucination Factors in Large Vision-Language Model via Causality Analysis

A. Details of Causal Implementations

A.1. Implementation of FGBG for Q Intervention

We design the *FGBG* prompting technique for S by asking the LVLM to describe the foreground main entities first and provide the description to prompt for the other background details. Specifically, we use the foreground prompt, "describe the foreground and ignore the background in the image" to obtain the foreground description A_f. Then, the prompt becomes "Given that the foreground is $[A_f]$, describe the other contents in the background."

B. Hallucination Analysis

We report the hallucination rate on the COCO validation subset [1] in terms of a single object (O_h), occurring objects (O_h^1, O_h^2), and the non-hallucinatory objects (O_n) in Table 1. The frequently hallucinated objects including a 'person', 'chair', and 'bottle' are the common main focus on a picture under the scope of the COCO categories [2], such as indoor, kitchen, food, and furniture. Our targeted non-hallucinatory objects are the ones that frequently occur with other hallucinatory objects. These objects include more scene-like nouns such as 'street' and 'ground'. Meanwhile, 'table' is commonly described in the context of other focusing objects as the objects in the indoor, kitchen, and food super categories are very likely located on a table. If we examine the responses containing more than one hallucination, the top five frequent words co-hallucinated are similar to the single-word hallucination. It is intriguing that 'handbag' is frequently hallucinated solely while not accompanied by other joint hallucination.

B.1. Unfaithful Visual Grounding and Irrelevant Tasks

A group of hallucination reduction research focuses on detecting objects and asks the model to refine the statement. This procedure involves setting up intermediate tasks and implementing object detectors. However, a lack of careful examination likely leads to neglected error propagation. We empirically find that the visual grounding techniques used in the previous works [4] show numerous detection failures, leading to unreliable hallucination revision. Following the approach in [4], we use an open-vocabulary object detector, GroundingDINO to detect the described objects in the image and then set up a discriminative task to ask the LVLM if the undetected objects exist in the image. We might expect the LVLM to be equipped with a stronger ability in binary classification problems. However, the results on the AMBER dataset shown in Table 2 are highly unsatisfactory. The low recall and F1 values demonstrate the issue of robustness for precise hallucination correction.

Therefore, we propose another avenue for hallucination reduction research, identifying the inducing non-hallucinatory objects or casual reasons instead of overly relying on auxiliary models. This novel direction mitigates the error propagation and explanability difficulties.

C. Additional Experiments

C.1. Direct Causal Effect Analysis

To distinguish the direct effect from overall effect, we additionally measure the direct causal effect (DCE). The DCE calculate the expected value $\mathbb{E}$ of the causal effect metric $\delta(P, P')$ over input X from an evaluation test after the intervention with specified mediator M by the equation:

$$DCE = \mathbb{E}_{x' \sim \mathbb{P}(X|M)}[\delta(P, P')], \tag{1}$$

Through controlling M, the direct effect can be distinguished from the total effect. For example, DCE can observe if the 'tree', an hallucinatory inducing word becomes less likely to drive other hallucinations after interventions. We discuss the DCE of our interventions on the hallucinatory inducing objects in Table 3. The CHAIR scores of 'water', 'sky', and 'beach' are greatly reduced by image pasting, and FGBG with InstructBLIP, among which the CHAIR score of 'sky' becomes 1.7% using image-pasting. Prominent CHAIR declines are also observable with mPLUG-Owl2. However, the word 'people' has been hallucinated using image-pasting, showing the effects of hallucination reduction are not uniform to each Z_o.

Table 3 also shows additional conditional analyses separately on the foreground (FG) and background (BG). The results using InstructBLIP have eminent differences between FG and BG with nearly zero hallucination occur-

	words	count	co-occurrence	count	induce	count
Instructblip	bottle	48	person,O_h^2	87	table	10
	person	46	people,O_h^2	69	street	3
	chair	42	bottle,O_h^2	65	ground	2
	car	38	chair,O_h^2	63	people	2
	cup	37	cup,O_h^2	56	plate	2
	words	count	co-occurrence	count	induce	count
mPlug-Owl2	person	64	person,O_h^2	113	people	25
	chair	53	chair,O_h^2	93	table	14
	people	49	people,O_h^2	86	ground	10
	handbag	41	cup,O_h^2	69	chair	9
	bottle	34	bowl,O_h^2	60	street	8

Table 1. This table shows the counts of the top 5 frequent (1) *single-hallucinatory* words O_h, (2) *co-occurring hallucinatory* words (O_h^1, O_h^2) (3) *Hallucinatory-inducing* words O_n. O_n is frequently associated with other hallucinatory words O_h, i.e., $O_n \rightarrow O_h$.

Instructblip Question	acc	precision	recall	f1
Instructblip	80.8	64.3	24.4	35.3
mPLUG-Owl2	81.2	74.1	19.5	30.8
mPLUG-Owl2 Question	**acc**	**precision**	**recall**	**f1**
Instructblip	77.5	75.7	35.9	48.7
mPLUG-Owl2	75.2	83.4	20.8	33.3

Table 2. Results of binary discriminative tasks described in §B.1. The questions generated based on the response of an LVLM are denoted with 'LVLM Question'. The results present responses to these questions using different LVLMs.

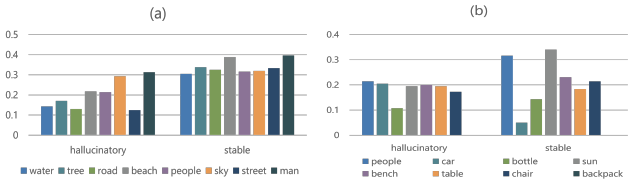

Figure 1. Retrieval score results using mPLUG-Owl2 on AMBER dataset for (a) non-hallucinatory words (O_{nh}) commonly co-occurring with O_h and (b) hallucinatory words (O_h).

ring with FG while still a few in BG. mPLUG-Owl2 generates hallucinations in a more even manner. However, it is intriguing that 'road' leads to 8.3 and 9.3 CHAIR scores with FG using InstructBLIP and mPLUG-Owl2, respectively. These analysis findings illuminate the hallucination pattern of an LVLM might be different due to the original training sets.

C.2. Foreground Prompt Generation Is Robust

Our proposed fore-background prompt strategy is a two-stage intervention forming a mediator to the causal graph. The underlying concept is similar to the chain-of-thoughts approach. We condense the chain into only two steps depending on the stability of the first step, foreground generation. In this way, we can estimate the expected value simply via a single inference.

We regenerate the foreground prompt three times and measure the hallucinatory scores with the results.

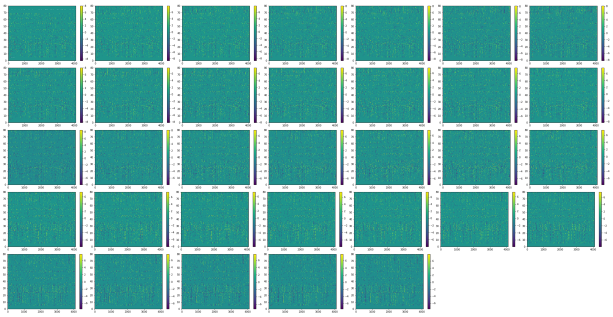

Figure 2. Embedding saliency map of the heads in mPLUG-Owl2. Most embedding saliency lie in similar dimensions.

C.3. Embedding Analyses

Our embedding analyses include two experiments, the visualization and retrieval safe scores. First, we extend the visualization results in the main manuscript by examining the detailed heads and timestamps without performing average pooling. Meanwhile, we present the full conditional results with the saliency map given that specific objects are mentioned in the response. Second, the retrieval safe scores demonstrate a group tendency of an image embedding locating closely with the non-hallucinatory images.

C.3.1 Visualization of Different Heads

We visualize the saliency maps of the different heads in mPLUG-Owl2 as shown in Figure 2. The resulting fig-

InstructBLIP	tree	water	sky	beach	road	mPLUG_Owl2	water	tree	road	beach	people
Baseline	30.6	21.6	23.8	20.5	33.3	Baseline	13.7	13.8	18.4	16.1	12.9
Image pasting	12.6	6.1	1.7	3.7	15.5	Image pasting	6.2	3.1	3.6	2.4	25
Stop	24.8	13.1	12.2	14.5	29.5	Stop	11.2	6.5	10.3	9.6	10
FGBG	15.1	9.1	11	8.7	25.3	FGBG	5.6	4.8	6.8	6.5	6.5
FG	1.9	0	0	0	8.3	FG	6.4	7.4	9.3	5.1	11.8
BG	17.5	11	14.1	10.7	27.6	BG	4.7	3.5	7.7	7.9	6.7

Table 3. CHAIR scores of text intervention: the 5 most common inducing non-hallucinatory words are investigated.

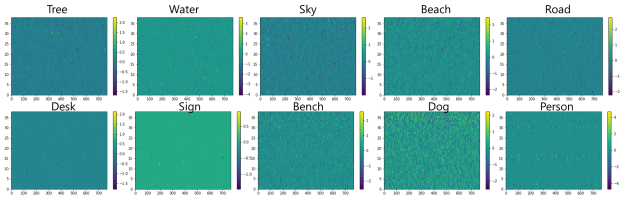

Figure 3. Embedding Saliency with timestamps in rows and dimensions in columns using InstructBLIP given a non-hallucinatory word commonly co-occurring with hallucinatory words in the response.

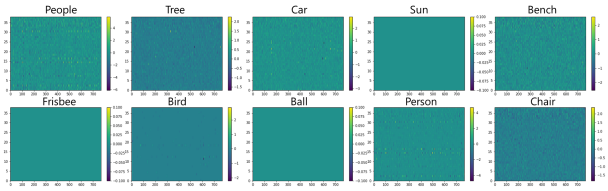

Figure 4. Embedding Saliency using InstructBLIP given commonly hallucinatory words in the response

ure indicates the significantly different values between the hallucinatory and non-hallucinatory objects. We use these saliency maps for embedding intervention. Therefore, we attempt to investigate the heads of the network for the salient dimensions. Consequently, most of the heads manifest with highly similar patterns that only very few dimensions on the sequence obtain significant values. This phenomenon implies the hallucinatory-related factors might exist in a few dimensions rather than widespread sitting in the whole dimensional space. This might correspond to extended research on seeking function vector [3] for the LLM, specific dimensions determine the tasks and properties of model outputs. We anticipate potential research delving into embedding editing to alter model behaviors without model retraining.

C.3.2 Visualization of Conditional Words

We illustrate the embedding saliency map under different conditions including the description with hallucinated objects generated (Figure 4) and the description with non-hallucinatory objects that are commonly generated accompanied by hallucinatory objects (Figure 3).

We observe that most non-hallucinatory objects bring about more salient dimensions that can differ between hallucinatory (X_h) and non-hallucinatory (X_n) images in terms of the embedding dimensions. For example, the 'sky', 'beach', and 'dogs' obtain high saliency in a few dimensions. On the contrary, 'frisbee' and 'sun' are hallucinatory examples showing limited saliency dimensions when ignoring those objects simultaneously common for non-hallucinatory words. The underlying reason might stem from the strong association of the 'frisbee' to a specific non-hallucinatory object. However, this object, e.g., 'tree', can be narrated along with a variety of other hallucinatory objects. The relation is naturally formed as a directed graph that perplexs the understanding of LVLM hallucination behaviors.

C.3.3 Retrieval Safe Scores

Following the experiments in §5.3 in the main manuscript, we regard the retrieval safe score to represent how likely a group of embeddings lying near the samples has seldom been hallucinated. Here, we report the results in Figure 1 for mPLUG-Owl2 given conditions on specific words generated in the response. If a word is a non-hallucinatory object (O_n) but frequently co-occurs with hallucinatory objects, we find that the cases in the hallucinated group X_h obtain lower retrieval safe scores than the stable group X_{stable}. However, a frequently hallucinatory object gives rise to the exception such as 'car'. The phenomenon might originate from the properties of non-hallucinatory objects that can be associated with various contexts. The contexts related to hallucinatory objects are different in contrast to other underlying context manifestations in the embeddings. The hallucinatory objects are usually generated under specific contexts. Even though this object is not hallucinated in a response, the context remains similar and thus hard to differentiate in the embedding space. The findings are consistent with the results of InstructBLIP described in the main manuscript.

(a_n, b_h)	All pairs	(court, ball)	(racket, ball)	(tennis, ball)	(sign, car)	(street, car)	(bench,people)	(train,people)	(shore,people)
InstructBLIP	0.173	0.455	0.455	0.385	0.33	0.308	0.243	0.208	0.2
Object removal	**0.068**	**0.364**	**0.364**	**0.308**	**0.24**	**0.192**	**0.146**	**0.041**	**0.04**
(a_n, b_h)	All pairs	(court, ball)	(racket, ball)	(umbrella, people)	(bench, people)	(tennis, ball)	(toilet,sink)	(house,person)	(hydrant, people)
mPLUG-Owl2	0.057	0.400	0.364	0.286	0.250	0.286	**0.095**	**0.067**	0.125
Object removal	**0.045**	**0.200**	**0.182**	**0.238**	**0.200**	**0.143**	0.238	0.133	**0.0625**

Table 4. The conditional object removal using InstructBLIP and mPLUG-Owl2 on AMBER. a_n denotes the hallucinatory-inducing object and b_h is the corresponding induced hallucinatory object. The $P(b_h \mid a_n, b_h)$ decreases when doing conditional object removal.

C.3.4 Conditional Object Removal Mitigate the Induced Hallucination

The object removal intervention does not achieve better results in overall CHAIR and HAL scores as shown in the main manuscript. We observe the result and find that object removal sometimes introduces other confounding factors, such as alterations to foreground-background semantic structures, perturbations induced by the limited ability of the inpainting technique, and erroneous removals. This results in unsatisfactory and inconsistent outcomes in hallucination reduction. To mitigate the effects of these confounding factors and focus on measuring the causal effect of the hallucinatory-inducing object O_n on the corresponding induced hallucinatory object O_h, we experiment with object removal conditional on the pair of O_n and O_h. We use the conditional probability $P(b_i \mid a_i, b_i)$ for $a_i \in O_n$ and $b_i \in O_h$. This examines whether the intervention of removing the hallucinatory-inducing object reduces the occurrence of its corresponding hallucinatory object. Table 4 shows the InstructBLIP and mPLUG-OWl2 result of the overall pairs and some pairs with the highest inducing hallucination rate on AMBER. When conditional on the hallucinatory-inducing and hallucinatory objects, the simple removal of the inducer a_n can significantly prevent inducing its corresponding hallucinated b_h.

C.3.5 Additional Implementations Details

We run all of our experiments on one NVIDIA GPU A6000 and one NVIDIA A40.

D. Case Analysis

We demonstrate some of the cases including the single-hallucinatory cases, co-occurring hallucinatory cases, and hallucinatory-inducing cases using InstructBLIP and mPLUG-Owl2 on the AMBER dataset as discussed in §3.2 of the main manuscript.

Image pasting (+)

"In the image, three people are walking through a lush green field, surrounded by tall grass and trees. One person is carrying a backpack, while the other two are walking together. A rabbit can be seen in the background, adding a playful touch to the scene. The group is likely enjoying a leisurely walk in the countryside, taking in the natural surroundings and enjoying each other's company."

Fore-background (-)

"a group of people walking through a grassy field. A group of three people is walking through a lush green field, surrounded by tall grass and trees. They appear to be enjoying the outdoors and taking a leisurely stroll together. One person is carrying a frisbee, suggesting that they may engage in a game or activity during their walk."

Stopping word (+)

"Instead, a group of three people are walking through a lush green field, enjoying the outdoors and each other's company. They appear to be having a good time as they stroll through the grassy area, possibly on their way to a picnic or other outdoor activity."

Figure 5. Case study for the hallucinatory 'frisbee'

Image pasting (+)

"a person riding a horse in the mountains, with snow-capped mountains in the background"

Fore-background (+)

"a man riding a horse in the mountains. A person is riding a horse on a grassy plain, surrounded by snow-capped mountains in the background. The horse is positioned towards the left side of the image, and the person appears to be enjoying the scenic view. There are several clouds visible in the sky, adding to the serene atmosphere of the scene."

Stopping word (-)

"In the scene, a person is riding a horse on a grassy plain, surrounded by snow-capped mountains in the background. The person is wearing a cowboy hat and appears to be enjoying the scenic view. The landscape is vast and open, with no trees or other vegetation visible in the immediate vicinity. "

Figure 6. Case study for single-hallucinatory 'tree' using InstructBLIP

Image pasting (+)

"The image depicts a sandy beach next to the ocean. There are several people in the scene, some of whom are walking along the beach, while others are further away from the water's edge. In the background, there is an aerial view of the beach and the ocean, showcasing the natural beauty of the location. The image captures the serene atmosphere of the beach."

Fore-background (-)

"an aerial view of a sandy beach next to the ocean. The image captures an aerial view of a sandy beach with two people walking along the water's edge. One person is closer to the camera, while the other is further away. The beach is surrounded by the ocean, with waves gently lapping against the shore. There is also a kite flying in the sky above the beach, adding to the serene atmosphere of the scene."

Stopping word (-)

"Instead, the image features an aerial view of a sandy beach with a single person walking along the edge of the water. There is no kite, sky, sound, sun, or surfboard in the scene."

Figure 7. Case study for 'water' induces the hallucinatory 'sky' using InstructBLIP

Image pasting (+)

"The image features an airplane flying in the sky, with a white rabbit sitting on top of the plane's wing. The rabbit appears to be perched on the edge of the wing, looking out towards the viewer. There are several clouds visible in the sky, adding to the overall atmosphere of the scene."

Fore-background (+)

"an airplane flying through a cloudy blue sky. A large airplane is flying through a cloudy blue sky, with its underbelly visible. The plane appears to be in the process of landing or taking off, as its landing gear can be seen at the bottom of the image. There are several clouds visible in the sky, adding to the dramatic atmosphere of the scene."

Stopping word (-)

"Instead, it features an airplane flying through a cloudy blue sky. The airplane appears to be in the process of landing, with its landing gear visible as it descends towards the ground."

Figure 8. Case study for 'sky' induces the hallucinatory 'ground' using InstructBLIP

Image pasting (+)

"The image depicts a small bathroom with a toilet. The toilet is situated in the middle of the room. There are two toilet paper rolls placed near the toilet, one on the right side and the other on the left side of the room. Additionally, there is a trash can located on the right side of the room, near the toilet paper roll on the right."

Fore-background (-)

"a small bathroom with a toilet and a trash can. The image features a small bathroom with a toilet and a trash can. The toilet is positioned in the corner of the room, while the trash can is placed next to the toilet. The bathroom has a tiled floor, which provides a clean and hygienic environment for the toilet and trash can."

Stopping word (-)

"The image shows a small bathroom with a toilet and a trash can. The toilet is located in the corner of the room, and the trash can is placed next to the toilet. There is no cup or sponge visible in the image."

Figure 9. Case study for 'paper rolls', and co-occurring hallucinatory 'cup' and 'sponge'

Image pasting (+)

"The image captures a thrilling moment of a motorcycle racer making a sharp turn on a winding road. The rider is skillfully navigating the curve, with his motorcycle leaning to the side.
In the background, there are several other people watching the motorcycle race, with some standing closer to the racer and others further away. The spectators appear to be engaged in the event, cheering on the racer as he makes his way around the corner.</s>"

Fore-background (+)

"In the foreground, there is a person riding a motorcycle down a street. The motorcycle is positioned towards the left side of the image, and the person is wearing a helmet while riding it. In the background of the image, there is a grassy area, which is located to the right of the person riding the motorcycle."

Original

"The image captures a thrilling moment of a motorcycle racer riding a motorcycle down a street. The rider is wearing a helmet and is fully focused on the road ahead. The motorcycle is positioned in the center of the scene, with the racer skillfully navigating the turns and curves of the street.

The surrounding area includes a few cars parked on the side of the street, a truck located further back, and a bench placed on the right side of the scene. The street is surrounded by grass, adding a touch of nature to the urban environment."

Figure 10. Case study for single-hallucinatory 'bench'

Image 3

Image pasting (+)

"In the image, a young boy is running across a grassy field with a white rabbit in the background. The boy is wearing a striped shirt and appears to be enjoying his time in the outdoors. The rabbit is positioned towards the right side of the image, adding a playful touch to the scene."

Fore-background (-)

"a toddler playing with a frisbee in a grassy field. In the image, a young boy is running across a grassy field, holding a frisbee in his hand. He appears to be enjoying himself as he runs and throws the frisbee around. There are several other people in the background, but they are not actively participating in the frisbee game."

Original

"In the image, a young boy is running through a lush green field, holding a frisbee in his hand. He appears to be enjoying himself as he runs and throws the frisbee back and forth. There are several other people in the scene, but they are not actively participating in the frisbee game. The grassy field is filled with various patches of green, creating a vibrant and lively setting for the young boy's playtime."

Figure 11. Case study for hallucinatory 'frisbee'

Image 136

Image pasting (+)

“In this image, the background features a clear blue sky and a vast expanse of sand dunes stretching as far as the eye can see.”

Fore-background (+)

“a person standing on top of a sand dune. A lone person is standing on top of a sand dune in the desert, looking out towards the horizon. The person is wearing a blue shirt and appears to be enjoying the scenery. In the background, there is a clear blue sky with a few clouds visible. The sand dune is quite large, stretching across the majority of the image."

Original

“In this image, a person is standing on top of a large sand dune, looking out towards the horizon. There are two people visible in the scene, one standing at the top of the dune and the other near the base of the dune. They appear to be enjoying the scenery and taking in the vast expanse of the desert landscape. A kite can be seen flying high up in the sky, adding to the serene atmosphere of the scene.”

Figure 12. Case study for hallucinatory ‘kite’ induced by ‘sky’ using InstructBLIP

Image 88

Image pasting (+)

“A woman is standing next to her bicycle, looking out over the ocean in the foreground. She appears to be enjoying the scenic view and taking in the beauty of the natural surroundings.”

Fore-background (+)

“a woman standing next to her bicycle, looking out at the ocean. A woman is standing next to a bicycle, looking out over the ocean. She is wearing a hat and appears to be enjoying the scenery. There is also a sign nearby, possibly indicating the location or providing information about the area."

Original

“The image features a woman standing next to a bicycle, looking out over the ocean. She is wearing a hat and appears to be enjoying the scenic view. The bicycle is positioned near the edge of the cliff, giving the woman a panoramic view of the ocean and the surrounding landscape. There are several other objects in the scene, including a bench located further away from the woman and the bicycle. The bench can be seen on the left side of the image, while the bicycle is situated closer to the edge of the cliff.”

Figure 13. Case study for hallucinatory ‘bench’ using InstructBLIP

Image 88

Image pasting (+)

“In the image, a young girl is standing on a beach and drinking from a green bottle. She is wearing a blue bathing suit and appears to be enjoying her time at the beach.

Fore-background (-)

“a young girl drinking from a bottle on the beach. A young girl is standing on a sandy beach, drinking from a bottle of beer. She is wearing a colorful swimsuit and appears to be enjoying her time at the beach. The bottle of beer is held close to her mouth as she takes a sip. In the background, there is a red kite flying in the sky, adding a playful touch to the scene."

Original

“The image features a young girl standing on a beach, drinking from a bottle. She is holding the bottle close to her mouth and appears to be enjoying her beverage. The bottle is positioned in the center of the scene, and the girl's head is slightly tilted towards the bottle as she takes a sip. In the background, there is a kite flying in the sky, adding a playful touch to the scene.”

Figure 14. Case study for hallucinatory ‘kite’ using InstructBLIP

Image 50

Image pasting (+)

“The image features a man standing next to a red car, possibly a Nissan, which is stuck in the snow. The man appears to be assisting the car in getting out of the snowy situation. </s>”

Fore-background (+)

“In the foreground, there is a man wearing black clothing, including a black hat, standing next to a red car. He appears to be walking or standing near the car, possibly in the snow. In the background, there is a snowy field with a wooden fence, adding to the wintry atmosphere of the scene."

Original

“The image shows a man standing next to a red car that is parked on a snow-covered road. The man is wearing a black jacket and appears to be getting out of the car. The car is parked on the side of the road, and there is another car further back in the scene.

In the background, there are two benches placed on the side of the road, possibly for resting or waiting for transportation. Additionally, there are two backpacks, one near the man and the other further away, possibly belonging to the man or other people in the area.”

Figure 15. Case study for hallucinatory ‘bench’ using mPLUG-Owl2

Image 415

Image pasting (-)

“The image features a man sitting in a chair on a lawn, enjoying a view of a lake. In front of him, a small white dog is on a leash, looking at the water. A white rabbit is also present in the scene, standing close to the dog.

In the background, there are two frisbees, one near the man and the other closer to the water. A cup can be seen placed on the lawn, possibly belonging to the man or someone else in the scene. The overall atmosphere is relaxed and peaceful, with the man and the animals sharing the serene lake view.</s>”

Fore-background (+)

“In the foreground, there is a man sitting in a chair on the grass. He is accompanied by a small white and gray dog, which is on a leash. The dog appears to be enjoying the outdoor setting, and the man seems to be relaxing and enjoying the company of his pet. In the background of the image, there is a lake with a boat floating on the water. The man and his dog are sitting in front of the lake, creating a serene and picturesque scene."

Original

“The image features a man sitting in a chair, enjoying a leisurely day by the water. He is accompanied by a small white and gray dog, which is tied to a rope and appears to be playing fetch with a frisbee. The dog is positioned closer to the water, while the man is seated in the chair on the grass.
In the background, there are two other people, one standing and the other sitting, possibly enjoying the scenery or engaging in their own activities. A car can also be seen in the background, parked near the water.”

Figure 16. Case study for hallucinatory ‘frisbee’ using mPLUG-Owl2

References

[1] 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. *arXiv preprint arXiv:2311.17911*, 2023. 1

[2] 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. 1

[3] Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron C Wallace, and David Bau. Function vectors in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[4] 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. *arXiv preprint arXiv:2311.13614*, 2023. 1