Supplementary Materials: Exploiting Semantic Reconstruction to Mitigate Hallucinations in Vision-Language Models

Minchan Kim[1†], Minyeong Kim[1†], Junik Bae[1†], Suhwan Choi[1], Sungkyung Kim[1], and Buru Chang[2*]

[1] Seoul National University
[2] Sogang University
{kjkk0502,kmy17518,heatz123,milkclouds,sk0428}@snu.ac.kr
buru@sogang.ac.kr

A Different Choices of Text-to-Image Models

ESREAL employs a text-to-image (T2I) model, specifically the 4-step SDXL Turbo model, to reconstruct images from model-generated captions. We explore how different T2I models impact ESREAL's performance. Specifically, we implement ESREAL on InstructBLIP FlanT5-XL (3B) using two different T2I models: the 1-step SDXL Turbo, which has inferior text-image alignment, and the 8-step Hyper-SDXL, which offers superior text-image alignment.

Table 1: T2I model comparison results.

T2I Models	Win Rate	FaithScore (↑)	GPT-4V (↓)
SDXL-Turbo (1 Step)	0.76	0.7484	2.23
SDXL-Turbo (4 Steps)	0.79	0.7834	1.49
Hyper-SDXL (8 Steps)	0.80	0.8141	1.32
DALLE-3	0.82	-	-

Table 1 illustrates that using a model with improved text-image alignment results in more stable rewards, as evidenced by increased win rates (see Section 5.2). These stable rewards significantly boost ESREAL's performance by reducing hallucinations, as demonstrated by higher evaluation scores on both the sentence-level FaithScore $\hat{f}_{\mathrm{s}}$ and the GPT-4V-aided evaluation (see Section 4.2).

Furthermore, we analyze the win rate of rewards with DALLE-3. Although the API-based image generation restricts full training due to cost and time constraints, the improved win rates support our findings that a robust T2I model yields stable rewards.

These findings suggest that ESREAL will soon benefit from more advanced image generators, currently constrained by speed and cost, to further minimize

† Equal contribution.
* Corresponding author.

hallucinations. The rapid advancement in the image generation field highlights ESREAL's potential to significantly improve its performance and reliability.

B Rationale for Semantic Reconstruction

When humans identify hallucinations within a generated caption, a pivotal step would involve comparing segments of the input image with the corresponding parts of the generated caption. For instance, to determine whether the statement 'A small white dog is sleeping.' contains hallucinations, one would first seek to locate the dog in the image, and then compare its features with the descriptions in the statement.

An ideal hallucination detection scheme would replicate this process. However, despite the technological advancements, the accurately binding portions of text to corresponding object regions in an image remains a challenging endeavor due to a difference in how objects are encapsulated in visual and textual data. Identifying which portion of the image corresponds to a specific object is relatively straight forward, because locality is clear in visual data. Typically, the pixels forming an object are confined to a specific region, allowing straightforward encapsulation, such as within a rectangular bounding box.

In contrast, in a paragraph, defining segments relevant to a specific object is considerably more complex. Unlike in images, the concept of object locality is not as pronounced in textual data. Descriptions of objects may be dispersed throughout the paragraph, making the search for segments of text pertinent to a specific object a difficult task. Even if the search is successful, another demanding aspect is reorganizing the segments to maintain the original semantic context. The segments of text within a paragraph are intricately connected to the surrounding phrases. Consequently, simple concatenation of the located object descriptions does not guarantee accurate interpretation, as their meaning can alter significantly without surrounding context. This necessitates a careful reorganization of the text to maintain the intended meaning, presenting another complex language task.

To surmount these difficulties, we propose a innovative solution: translating the text generated by the model to the visual domain. By converting model generated paragraphs into images, binding regions pertinent to a specific object becomes a easier task. Then, a comparative analysis on these aligned regions reveals hallucinations. Our approach in identifying hallucinations through semantic reconstruction and training models to mitigate hallucinations proves to be an effective approach in tackling hallucinations. A significant benefit of our method is that it is entirely unsupervised, eliminating the need for extra data beyond the images themselves.

C Training Details

C.1 Dataset Statistics

The Stanford Image Paragraph Dataset [2] contains a total of 19561 images with 14579 train images, 2490 validation images, and 2492 test images. The length of the corresponding detailed captions are on average 69.69 BERT tokens, 60.88 words, and 309.71 characters.

C.2 Training Hyperparameters

Table 2: Training hyperparameters.

HPs	InstructBLIP	LLaVA	mPLUG-Owl2
Training steps	800	800	800
Batch size	64	16	16
Learning rate	1e-5	1e-5	1e-5

Note: All models use AdamW optimizer with $\epsilon = 1e-6$, $\beta = (0.9, 0.95)$, and weight decay of $1e-6$.

We describe the hyperparameters for ESREAL as follows: The sequence length is configured to be 256, with a total of 800 training steps. The batch size is set at 64 for InstructBLIP, while for LLaVA and mPLUG-Owl2, it is 16. The optimization is performed using the AdamW Optimizer, which has betas set to (0.9, 0.95), epsilon at 1.0e-8, and a weight decay of 1.0e-6. Additionally, a cosine annealing scheduler is employed. For Proximal Policy Optimization (PPO), we use 4 epochs, with gamma set to 1 and lambda at 0.95. The clipping range for PPO is 0.2, similar to the value function clipping range, which is also 0.2. The coefficient for the value function is set at 1, and the clipping range for rewards is established at 10. The reward function is weighted with alpha at 0.8 and beta at 0.001.

C.3 Trainable Components in VLMs

Due to the substantial number of parameters in contemporary VLMs, it is standard practice to keep some components of VLMs frozen while allowing others to be trainable [3, 5]. We also adopt this approach when applying ESREAL on LLaVA 1.5, InstructBLIP, and mPLUG-Owl2. In determining the trainable components, we generally comply with the the training strategies specific to each VLM, with an exception to InstructBLIP, where we additionally train the LLM. For all fine-tuning processes, we employ Low-Rank Adaptation for large language models (LoRA) [1]. The trainable modules for each VLM are detailed in Table 3.

Table 3: Trainable VLM components.

Components	LLaVA	InstructBLIP	mPLUG-Owl2
Visual encoder	x	x	o
Modality connection module	o	o	o
LLM	o	o	o

Note: ‘o’ denotes trainable components, and ‘x’ denotes frozen components.

C.4 Optimization of Reward Model

ESREAL employs several large-scale models in its reward mechanism, necessitating efficient operation within acceptable memory and time constraints. Proximal Policy Optimization (PPO) offers the advantage of not requiring a differentiable reward model, allowing us to improve throughput and latency by utilizing a NVIDIA Triton Inference Server.

We deploy key components of the reward model —SDXL Turbo (reconstruction module), Grounding DINO (alignment module), and the CLIP encoder (scoring module)— on this server. The critical bottleneck of our training procedure, the reconstruction and alignment modules, are distributed across multiple GPUs. The optimized architecture and Triton Inference Server code can be found on our GitHub.

This optimization enables a single reward computation cycle, which calculates the aggregated penalty for four reconstructed images generated from one caption, to complete in 3.27 seconds on a single GPU.

C.5 Module-specific Details

Positional Tokens Positional tokens are tokens that denote spatial relationships. Specifically, we view [‘left’, ‘right’, ‘top’, ‘bottom’, ‘center’, ‘middle’, ‘above’, ‘below’, ‘inside’, ‘outside’, ‘front’, ‘behind’, ‘upward’, ‘downward’, ‘up’, ‘down’, ‘inward’, ‘outward’, ‘over’, ‘under’] as positional tokens.

Case Study In Figure 1, we demonstrate a case study of our alignment module and scoring module.

The alignment module first aligns the caption with the reconstructed image. As a result, the phrases - elderly man, guide dog, red traffic light, and yellow taxis - are matched to their corresponding regions in the reconstructed image. Then, the concatenated phrase ‘elderly man, guide dog, red traffic light, yellow taxis’ is aligned to the original image. As a result, only ‘yellow taxis’ remains unaligned while the other phrases find matching regions in the original image. Using the object phrases as an anchor, regions from the reconstructed image and the original image are aligned.

The scoring module produces hallucination penalties for non-existent objects, unfaithful attributes and inaccurate relationships. The unaligned ‘yellow taxis’

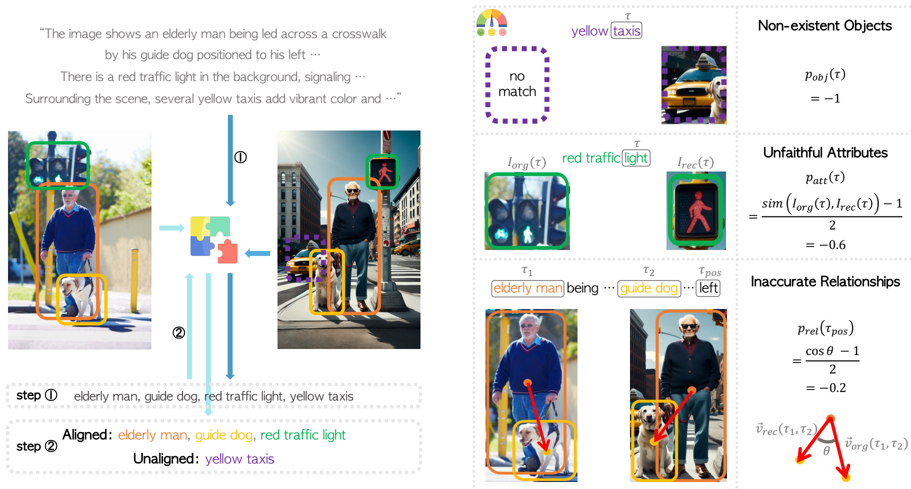


Fig. 1: Detailed illustration of the alignment module and the scoring module.

reveals a hallucinated object. Accordingly, a non-existent object penalty of -1 is allocated to 'taxis'. The aligned regions for 'red traffic light' are passed to the CLIP encoder and receive a low similarity score of -0.6, due to the misrepresentation of a green light as red. This unfaithful object attribute penalty is allocated to 'light'. The object phrases 'elderly man' and 'guide dog' are paired because they are linked within the same sentence through a positional token 'left'. Two vectors are constructed by connecting the center points of the aligned regions for 'elderly man' and 'guide dog' in the original and reconstructed image. The orientation of these two vectors, assessed by the cosine of their angle, receives a -0.2 penalty, indicating an inaccurate spatial relationship. The penalty is allocated to the positional token 'left'.

D ESREAL on COCO

We conduct our main experiments on the Stanford Image Paragraph Dataset [2]. The dataset offers the benefit of being able to evaluate ESREAL on standard captioning metrics due to its human-annotated captions. Therefore, we fine-tune VLMs with the training images of the dataset and evaluate their performance on its test split.

However, since ESREAL is an unsupervised method, one can utilize any image dataset, including those without annotated detailed captions. To showcase ESREAL's versatility across different datasets, we conduct experiments using the MS COCO 2014 dataset [4]. Specifically, we fine-tune InstructBLIP FlanT5-XL on 32,000 images sampled from the COCO 2014 Karpathy train split and evaluated its performance on the test split of the Stanford Image Paragraph Dataset. The evaluation of ESREAL on COCO aligns with the findings from the

Table 4: CHAIR evaluation results on COCO.

Model	Method	CHAIR		
		$CHAIR_s$ (↓)	$CHAIR_i$ (↓)	Coverage (↑)
InstructBLIP	Baseline	0.48	0.16	**0.36**
	ESREAL	**0.45**	**0.15**	**0.36**

Table 5: Image paragraph captioning task evaluation results on COCO.

Model	Method	CIDEr (↑)	ROUGE-L (↑)	BLEU (↑)			
				BLEU-1	BLEU-2	BLEU-3	BLEU-4
InstructBLIP	Baseline	5.99	**22.68**	26.79	13.68	6.76	3.51
	ESREAL	**8.31**	**22.68**	**30.43**	**15.48**	**7.69**	**4.06**

main experiment, demonstrating its ability to mitigate hallucinations without impairing the generative performance of VLMs.

D.1 Hallucination Evaluation

We report the CHAIR metric to evaluate the generated captions in terms of hallucinations. Table 4 shows that ESREAL effectively decreased both the sentence level and object instance level CHAIR scores while maintaining Coverage, indicating successful mitigation of hallucinations.

D.2 Captioning Evaluation

We also report the standard captioning metric to evaluate the quality of the generated captions. Table 5 shows that ESREAL improves captioning performance beyond the baseline InstructBLIP, notably increasing the CIDEr score.

E Qualitative Examples

We provide additional qualitative examples to demonstrate the effectiveness of ESREAL. Figures 2, 3, 4 show that ESREAL effectively mitigates hallucinations in captions across different types of hallucinations.

In figure 2, the caption generated by baseline LLaVA contains non-existent objects 'dining table visible in the background' and 'another chair placed further away'. There is also an incorrect attribute 'two chairs'. After fine-tuning with ESREAL, LLaVA is able to correctly identify the 'blue background' and 'a chair'.

Figure 3 shows an example of an inaccurate spatial relationship in the original caption, 'some sheep standing on top of it'. Post ESREAL, the caption accurately depicts the spatial relationship between the sheep and hay pile as 'some closer

The image features a man wearing a suit and tie, sitting in a chair and speaking into a microphone. He appears to be giving a speech or participating in a discussion. The man is wearing glasses and is focused on his speech. The scene takes place in a room with a dining table visible in the background. There are two chairs in the room, one occupied by the man speaking into the microphone and another chair placed further away.

The image features a man sitting in a chair and talking into a microphone. He is wearing a suit and glasses, and his mouth is open as he speaks. The man appears to be engaged in a conversation or giving a speech. The scene is set against a blue background.

Fig. 2: Captions generated by LLaVA before and after ESREAL.

The image features a large group of sheep gathered around a pile of hay in a field. There are at least 13 sheep in the scene, with some standing closer to the hay and others spread out around the field. The sheep are of various sizes, indicating a mix of adult and younger sheep. The hay pile is located in the center of the field, with some sheep standing on top of it, while others are surrounding it. The sheep appear to be enjoying their time together, grazing and socializing in the open field.

The image features a large herd of sheep gathered around a pile of hay in a field. The sheep are scattered around the hay, with some close to the hay and others further away. The field is filled with sheep, creating a lively scene. In addition to the sheep, there are a few sheep located in the background, adding to the overall scene. The field is filled with sheep, creating a vibrant and bustling atmosphere.

Fig. 3: Captions generated by LLaVA before and after ESREAL.

The image features a man riding in a horse-drawn carriage, with two horses pulling the wagon. The man is sitting in the back of the wagon, enjoying the ride. The horses are positioned on either side of the carriage, with one horse on the left and the other on the right.
In the background, there are several cars parked, indicating that the scene takes place in a parking lot or a similar area. The presence of the cars suggests that the horse-drawn carriage is a unique and nostalgic mode of transportation in this location.

The image features a man sitting in a wagon being pulled by two horses. The wagon is an old-fashioned, covered wagon with a canvas top, and it appears to be a stagecoach. The man is seated in the back of the wagon, enjoying the ride. The horses are positioned on either side of the wagon, with one horse on the left side and the other on the right side. They are pulling the wagon down a dirt road, creating a nostalgic and picturesque scene.

Fig. 4: Captions generated by LLaVA before and after ESREAL.

to the hay and others further away'. Additionally, it is able to capture the 'few sheep located in the background'.

The baseline caption in figure 4 contains a hallucinated object 'several cars'. LLaVA with ESREAL, however, does not fabricate non-existent cars anymore. Moreover, it is able to identify additional attributes of the wagon such as 'old-fashioned', 'covered with a canvas top' and the 'dirt road'.

F GPT Prompting Details

F.1 GPT-4V-aided evaluation

Figure 5 provides the prompt we used for GPT-4V-aided evaluation. Figure 6 illustrates an example output of GPT-4V-aided evaluation.

F.2 Generating Hallucinated Captions for Win Rate Analysis

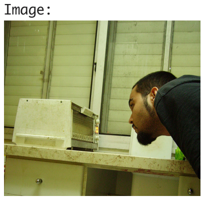
Figures 7, 8, 9 show the prompts we used to generate hallucinated captions containing non-existent objects, unfaithful object attributes, and inaccurate relationships, respectively. The hallucinated captions were used to calculate the Win Rate of rewards in the Stability Analysis section. Figure 10 provides an example of the hallucinated captions generated by GPT-4.

```
Given a single image and its accompanying captions, identify any discrepancies known as 'hallucinations' in
the captions relative to the actual content of the image. Your response should exclusively list these
hallucinations in JSON format, ensuring the sequence of the analyzed caption matches that of the reported
hallucination. Focus on hallucinations that are categorically incorrect according to the following
classifications:

1. Object Hallucinations: These occur when a caption mentions an object that is not present in the image.
For instance, if there is no carrot in the image but the caption says "a red apple and an orange carrot,"
this is an object hallucination.

2. Attribute Hallucinations: These happen when a caption describes an attribute of an object that does not
align with its appearance in the image. For example, describing an apple in the image as green when it is
actually red.

3. Relationship Hallucinations: These arise when a caption inaccurately describes the spatial or relational
dynamic between objects in the image. For example, if an apple is on a table but the caption says it is
underneath the table, it's a relationship hallucination.


Guidelines:

- Only report explicit hallucinations in JSON format. The output should directly correspond to the input
captions.

- Focus solely on hallucinations that are directly contradicted by the image content. Subtle inaccuracies
or generalizations (e.g., specifying "black horse" when the image shows a horse of an unspecified color, or
"three people" instead of "some people") do not constitute hallucinations under this task.


Example of expected output format:
## Input
Image: (An image which contains a man standing to the left of the black horse)
Caption: "A man in a cowboy hat is standing to the right of the white horse."
## Output
{"hall_object": ["cowboy hat"], "hall_attribute": ["white horse"], "hall_relationships": ["right"]}

## Input
Image: (An image which contains a man standing to the left of the black horse)
Caption: "A man is standing to the left of the yellow horse."
## Output {"hall_object": [], "hall_attribute": ["yellow horse"], "hall_relationships": []}

## Input
Image: (An image which contains a man standing to the left of the black horse)
Caption: "A man is standing to the left of the horse."
## Output {"hall_object": [], "hall_attribute": [], "hall_relationships": []}


# Task
Analyze the caption(s) provided below against the associated image and identify any hallucinations as
instructed.

## Input
Image:
```

Fig. 5: Prompt for GPT-4V-aided evaluation.

Input

Image:



Caption:

The image features a man with a beard, wearing a black shirt, standing in front of a counter. He appears to be staring intently at something on the counter, possibly a microwave. The microwave is located towards the right side of the counter, and the man is standing close to it.

The countertop is also adorned with a bottle, which is located near the right side of the scene. The overall atmosphere suggests that the man is focused on something on the counter, possibly preparing a meal or checking the microwave.

Output

'hall_object': ['bottle'], 'hall_attribute': [], 'hall_relationships': ['right']

Fig. 6: Example output of GPT-4V-aided evaluation.

```
You are an AI assistant who has strong reasoning abilities.

You will be provided with:
1. A paragraph description of an image.
2. A list of objects.

Your goal is:

To create a hallucinated paragraph by adding some of the objects in the list to the paragraph as
hallucinated objects.

Please follow the following rules:
1. Add the hallucinated objects to the paragraph in a way that makes sense.
2. Refrain from changing the paragraph as much as possible.
3. Do not add any object other than those given in the list.
4. Do not remove objects that exist.

Example 1)

Create a hallucinated paragraph by adding some of the objects in the list to the paragraph as hallucinated
objects. Do not remove objects that exist.

Paragraph: The bird is on a rock. The bird is brown and red. The bird is very small. The bird has very
jagged feet. The bird has black eyes.

Objects: a man

Result: There is a man. A bird is on a rock. The bird is brown and red. The bird is very small. The bird
has very jagged feet. The bird has black eyes.

Example 2)

Create a hallucinated paragraph by adding some of the objects in the list to the paragraph as hallucinated
objects. Do not remove objects that exist.

Paragraph: Three people are riding on horses on the beach on a gray day. The horse in the front is brown,
and there are two horses behind it. Behind the black horse there is a white horse. The person on the front
horse is wearing a gray jacket and a black helmet and the next person is in a dark blue jacket and helmet.
The final person is wearing a royal blue jacket and a black helmet.

Objects: sport car

Result: There is a sport car. Three people are riding on horses on the beach on a gray day. The horse in
the front is brown, and there are two horses behind it. Behind the black horse there is a white horse. The
person on the front horse is wearing a gray jacket and a black helmet and the next person is in a dark blue
jacket and helmet. The final person is wearing a royal blue jacket and a black helmet.

Now here is the actual task.

Create a hallucinated paragraph by adding some of the objects in the list to the paragraph as hallucinated
objects. Do not remove objects that exist.

Paragraph: {paragraph}

Objects: {objects}

Result:
```

Fig. 7: Prompt for generating hallucinated captions with non-existent objects.

```
You are an AI assistant who has strong reasoning abilities.

You will be provided with a paragraph description of an image.

Your goal is:

To create a hallucinated paragraph by modifying some properties of objects such as color, size, and shape.

Please follow the following rules:
1. You can only modify the properties of objects that are mentioned in the paragraph.
2. Only add hallucinations to the properties of objects that already exist in the paragraph.
3. Do not add objects that do not exist.
4. Do not remove objects that exist.
5. You should modify the properties of only one object in the paragraph.

Example 1)

Create a hallucinated paragraph by modifying some properties of objects. Do not add objects that do not
exist. Do not remove objects that exist.

Paragraph:

Two people stand on the sand holding a big C-kite. The C-kite is color red, orange and green. The men wear
black clothes. Shadows are cast on the ground. There are foot prints of a dog on the sand. In front of the
men, there is a body of water. Next to the body of water there are three people. Behind the body of water
there are small hills.

Result:

Two people stand on the sand holding a big C-kite. The C-kite is color red, orange and green. The men wear
white clothes. Shadows are cast on the ground. There are foot prints of a dog on the sand. In front of the
men, there is a body of water. Next to the body of water there are three people. Behind the body of water
there are small hills.

Example 2)

Create a hallucinated paragraph by modifying some properties of objects. Do not add objects that do not
exist. Do not remove objects that exist.

Paragraph:

Three people are riding on horses on the beach on a gray day. The horse in the front is brown, and there
are two horses behind it. Behind the black horse there is a white horse. The person on the front horse is
wearing a gray jacket and a black helmet and the next person is in a dark blue jacket and helmet. The final
person is wearing a royal blue jacket and a black helmet.

Result:

Three people are riding on horses on the beach on a gray day. The horse in the front is brown, and there
are two horses behind it. Behind the black horse there is a white horse. The person on the front horse is
wearing a gray jacket and a red helmet and the next person is in a dark blue jacket and helmet. The final
person is wearing a royal blue jacket and a black helmet.

Now here is the actual task.

Create a hallucinated paragraph by modifying some properties of objects. Do not add objects that do not
exist. Do not remove objects that exist.

Paragraph:

{paragraph}

Result:
```

Fig. 8: Prompt for generating hallucinated captions with unfaithful object attributes.

```
You are an AI assistant who has strong reasoning abilities.

You will be provided with a paragraph description of an image.

Your goal is:

To create a hallucinated paragraph by modifying some relationship between objects such as relative
positions between objects.

Please follow the following rules:
1. You can only modify the relationships between objects that are mentioned in the paragraph.
2. Only add hallucinations to the relationships between objects that already exist in the paragraph.
3. Do not add objects that do not exist.
4. Do not remove objects that exist.
5. You should modify the relationships between only one pair of objects in the paragraph.

Example 1)

Create a hallucinated paragraph by modifying some relationship between objects such as relative positions
between objects.

Paragraph:

Two people stand on the sand holding a big C-kite. The C-kite is color red, orange and green. The men wear
black clothes. Shadows are cast on the ground. There are foot prints of a dog on the sand. In front of the
men, there is a body of water. Next to the body of water there are three people. Behind the body of water
there are small hills.

Result:

Two people stand on the sand holding a big C-kite. The C-kite is color red, orange and green. The men wear
black clothes. Shadows are cast on the ground. There are foot prints of a dog on the sand. Behind the men,
there is a body of water. Next to the body of water there are three people. Behind the body of water there
are small hills.

Example 2)

Create a hallucinated paragraph by modifying some relationship between objects such as relative positions
between objects.

Paragraph:

Three people are riding on horses on the beach on a gray day. The horse in the front is brown, and there
are two horses behind it. Behind the black horse there is a white horse. The person on the front horse is
wearing a gray jacket and a black helmet and the next person is in a dark blue jacket and helmet. The final
person is wearing a royal blue jacket and a black helmet.

Result:

Three people are riding on horses on the beach on a gray day. The horse in the front is brown, and there
are two horses behind it. In front of the black horse there is a white horse. The person on the front horse
is wearing a gray jacket and a black helmet and the next person is in a dark blue jacket and helmet. The
final person is wearing a royal blue jacket and a black helmet.

Now here is the actual task.

Create a hallucinated paragraph by modifying some relationship between objects such as relative positions
between objects.

Paragraph:

{paragraph}

Result:
```

Fig. 9: Prompt for generating hallucinated captions with inaccurate relationships.

1. **Non-existent Object**

Create a hallucinated paragraph by modifying some properties of objects. Do not add objects that do not exist. Do not remove objects that exist.

Paragraph:

A person is skiing through the snow. There is loose snow all around them from him jumping. The person is wearing a yellow snow suit. The person is holding two ski poles in their hands.

Result:

There are signs. A person is skiing through the snow. There is loose snow all around them from him jumping. The person is wearing a yellow snow suit. The person is holding two ski poles in their hands.

2. **Unfaithful Object Attribute**

Create a hallucinated paragraph by modifying some properties of objects. Do not add objects that do not exist. Do not remove objects that exist.

Paragraph:

The man is taking a photo in the round mirror. He is bald. He is wearing an orange jacket. His camera is black. There is a train in the mirror too.

Result:

The man is taking a photo in the round mirror. He is bald. He is wearing an orange jacket. His camera is silver. There is a train in the mirror too.

3.**Incorrect Relationship**

Create a hallucinated paragraph by modifying some relationship between objects such as relative positions between objects.

Paragraph:

A woman in a blue tennis outfit stands on a green tennis court. She is swinging a blue tennis racket. There is a green tennis ball above her head.

Result:

A woman in a blue tennis outfit stands on a green tennis court. She is swinging a blue tennis racket. There is a green tennis ball below her head.

Fig. 10: Example output of generating hallucinated captions for Win Rate analysis.

References

1. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models (2021)
2. Krause, J., Johnson, J., Krishna, R., Fei-Fei, L.: A hierarchical approach for generating descriptive image paragraphs. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 317–325 (2017)
3. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
4. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
5. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=w0H2xGHlkw