



Appendix for Hallucination Augmented Contrastive Learning for Multimodal Large Language Model


1. Case Study


Results are shown in Figure 1. In the first example, the query prompts the models to describe the image. However, LLaVA, while correctly describing some content, produced hallucinatory information that was not present in the image, such as mentioning non-existent bicycles and other objects. MiniGPT-4 did not accurately describe the content of the picture. After applying our method to these models, we found that LLaVA-HACL and MiniGPT-4-HACL accurately answered the question. In the second example, we inquired whether certain non-existent objects were present in the picture, and both LLaVA and MiniGPT-4 incorrectly answered yes. After applying our method, these models correctly responded to the questions.


2. Generation of Hallucinative Captions


As shown in Table 1, we provide GPT-4 with ground truth captions for images, along with multiple in-context examples, guiding it to generate hallucinative captions based on the definition of hallucination and the provided examples.


Describe this image.  





A girl is riding down a smooth concrete surface in a parking lot. A bicycle can also be seen in the background, parked on the right side of the parking lot.  LLaVA


A little girl dressed in a pink t-shirt is riding a scooter on a smooth concrete surface, with other people providing a backdrop to this playful scene.  LLaVA-HACL


This picture shows a girl and a boy riding bicycles on the road, with a forest behind them.  MiniGPT-4


A girl is riding a scooter on the park road, with a boy and a forest behind her.  MiniGPT-4-HACL



Example #1


Is there a TV in the image?  





Yes, there is a TV in the image, located on a shelf above the bed.  LLaVA

No, there is not a TV in the image.  LLaVA-HACL

Is there a giraffe in the image?  



Yes, there is a giraffe in the image.  MiniGPT-4

No, there is a elephant in the image.  MiniGPT-4-HACL

Example #2

Figure 1. Example cases compared with MiniGPT-4, LLaVA and the variants equipped with our HACL.

System Prompt:

Hallucination in Large-scale Visual Language Models (LVLMs) refers to cases where these models generate descriptions introducing elements that are inconsistent with the content or completely absent from a provided image. These hallucinations can be coarse-grained, focusing on the mere existence of objects, or fine-grained, focusing on more specific attributes or characteristics such as quantity, properties, and locations. It's noteworthy that LVLMs often hallucinate about objects frequently present in visual instructions or within the actual image contents.

Your task is to revise a given caption to create a mirrored version that closely aligns with the original's content and length but incorporates elements of hallucination. Please craft two versions of this 'hallucinated' caption, each one representing either coarse-grained or fine-grained hallucinations. The first step involves identifying the objects involved and their associated attributes within the given caption. Subsequently, combine this insight with the details concerning hallucinations provided above to complete your task.

In-context Examples:

Input: A woman stands in the dining area at the table.

Output: A woman sitting in the classroom in front of the blackboard

====

Input: A room with chairs, a table, and a woman in it.

Output: A room with a fireplace, a computer and a man in it

====

Input: The large brown bear has a black nose.

Output: The cut dog owns a brown nose and blue eyes

====

Input: Bedroom scene with a bookcase, blue comforter and window.

Output: Bedroom scene with green bed, and a cupboard

====

Input: Many cars traveling on a busy road with exit signs overhead.

Output: Many cars traveling on a busy road and there is a plane flying in the sky.

====

Input: A couple of baseball player standing on a field.

Output: Two football players running on the grass.

====

Input: A man in a blue shirt is standing on a beach.

Output: A man in a red shirt wearing glasses is standing in a kitchen

====

Input: A woman is sitting on a bench in the park.

Output: A handsome man is standing on a bench in the park and watching the sky

====

Input: A woman is standing in front of a blue door.

Output: A woman is standing in front of a green car on the street

====

Input: A man in a blue shirt and jeans is riding a skateboard.

Output: A man in a purple shirt and shorts is riding a bicycle.

Table 1. One example used in in-context-learning to generate hallucinative captions.