Attention Never Lie: Visual Attention Defocus Reveals and Rectifies Hallucinations in MLLMs

Supplementary Material

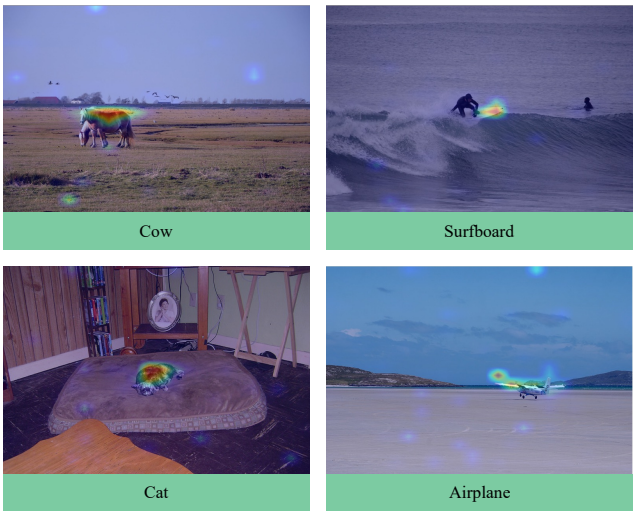


Figure A1. When MLLMs generate correct answers, their attention is relatively concentrated.

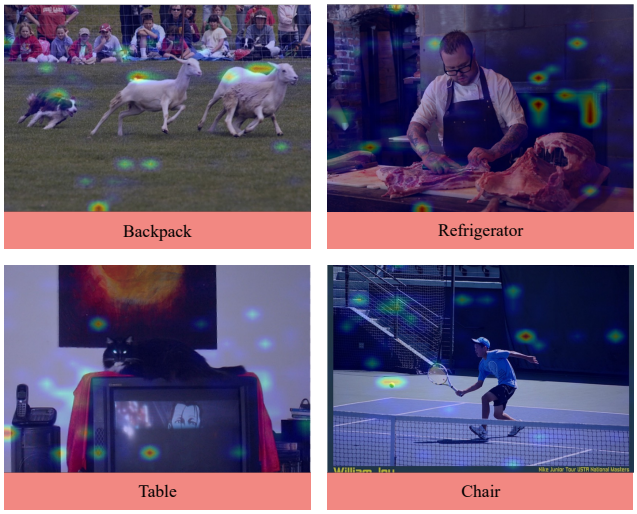


Figure A2. When hallucinations occur in MLLMs' answers, the phenomenon of attention defocus arises.

7. More attention visualization examples

In this section, we provide additional visualization results to illustrate the differences in attention when MLLMs answer questions. As shown in Fig. A1, when MLLMs generate correct answers, their attention is often concentrated on the correct regions. In contrast, when MLLMs produce hallucinatory responses, their attention tends to be highly scattered (as presented in Fig. A2), exhibiting a phenomenon of attention defocus.

8. Illustration of CHAIR's limitations and Prompt for GPT-4o

In this section, We highlight the shortcomings of using CHAIR as a hallucination evaluation metric. As shown in Fig. A3, when 'seat' refers to a passenger's seat rather than a physical object, this method may result in incorrect judgments. Additionally, there are issues with tokenization and unclear part-of-speech tagging.

Thus, to avoid these issues, we have selected GPT-4o (`gpt-4o-2024-05-13`) to detect hallucinations in captions. Tab. A1 presents the prompt template we used for hallucination detection with GPT-4o.

9. MLP-guided Search

9.1. Algorithmic flowchart of our method

To provide a detailed illustration of our method, we offer an example of combining it with Beam Search, whose algorithmic flowchart is presented in Alg. 1.

9.2. MLP Model Architecture

In our experiments, we designed and trained a Multi-Layer Perceptron (MLP) model to detect differences between hallucinated samples and clean samples. Specifically, the input to the model is a processed feature vector with a dimension of 576×32×32. The structure of the MLP model consists of four fully connected layers. Each of the first three layers includes a linear layer, a BatchNorm regularization layer, a ReLU activation function, and a Dropout layer (with a dropout rate of 0.5). The final layer is a linear layer used to output the binary classification results.

During training, we used the Binary Cross-Entropy Loss function (BCEWithLogitsLoss) to optimize the model parameters and employed the Adam optimizer (with an initial learning rate of 0.001) for gradient updates. To improve the generalization ability of the model, we added BatchNorm and Dropout operations after each linear layer. Additionally, a learning rate scheduler (ReduceLROnPlateau) was utilized to dynamically adjust the learning rate when the validation loss did not decrease significantly.

Caption: The image features a person wearing a green hat and a dog sitting in the passenger **seat** of a truck. The person is driving the truck, and the dog appears to be enjoying the ride. The truck is parked, and the driver is likely taking a break or waiting for someone.\n\nIn the scene, there are also two other people visible, one standing near the truck and another person further away. Additionally, there are two balloons in the scene.

CHAIR detects hallucinated words: **seat, chair**
Ground truth words: truck, person, dog, car

Caption: The image features a group of four teddy bears sitting on a blanket, arranged as if they are having a picnic. The teddy **bears** are positioned in a row, with each one holding a plate in their hands. The plates are placed in front of the bears, creating a charming scene.\n\nIn addition to the teddy bears and plates, there are two bowls on the blanket, one near the left side and the other near

CHAIR detects hallucinated words: **bear**
Ground truth words: truck, teddy bear, bowl

Caption: The image shows a snowboarder wearing goggles and an **orange** sweater, flying through the air over a snow-covered slope. The snowboarder is performing an impressive jump, soaring above the snow-covered surroundings.\n\nThe scene portrays that the snow-covered mountain is inhabited, as a building can be spotted nearby. Interestingly, there is a cat puzzled by the snowboarder's antics, curiously observing the scene from below. Multiple snowboards can also be observed in the image, although they are not in use.",

CHAIR detects hallucinated words: **orange,** cat
Ground truth words: snowboard, person

Caption: The image depicts a street scene with a red and white \"No **Motorcycles**\" sign prominently displayed on the side of the road. The sign is positioned in front of a building, possibly a shop or a restaurant. \n\nIn addition to the \"No **Motorcycles**\" sign, there is a \"No **Bicycles**\" sign further down the street, indicating that both **motorcycles** and **bicycles** are not allowed in the area. A car can be seen parked on the side of the street, and a person is standing nearby, possibly observing the signs or waiting to cross the street

CHAIR detects hallucinated words: **motorcycle, bicycle**
Ground truth words: car, person

Figure A3. Some error examples of hallucination annotations using the chair method [37].

As a multimodal analysis expert, You will be provided with an image and a description of that image. Your task is to determine if the description contains any information that cannot be directly observed in the image (hallucinations). It should only be marked as containing hallucinations if it describes elements that are definitively not present in the image. Therefore, you should carefully observe the image before detecting hallucinations: **Image Parsing** - Object detection: List all discernible objects with object and their position,color and quantity - Spatial relationships: Map relative positions - Text elements: Extract legible text.

Important Guidelines:

1. Focus solely on what is visually present in the image
2. Only Consider spatial relationships, colors, actions, quantities, and object properties
3. Be conservative in your judgment - only mark something as a hallucination if you are certain it's not in the image
4. Do not mark the words which is unclear in image as hallucinations

Now, analyze the following description:
Description: {description}

Instructions:

1. Split the description into words
2. Identify each word that represents a hallucination
3. Briefly explain why each word is a hallucination

Response Format:

Hallucinated words:

- **Word:** [only word]
- **Reason:** [brief explanation with specific reference to what is actually shown in the image]

Note: Only include words that are definitely hallucinations

Table A1. Text prompt template for hallucination detection using GPT-4o.

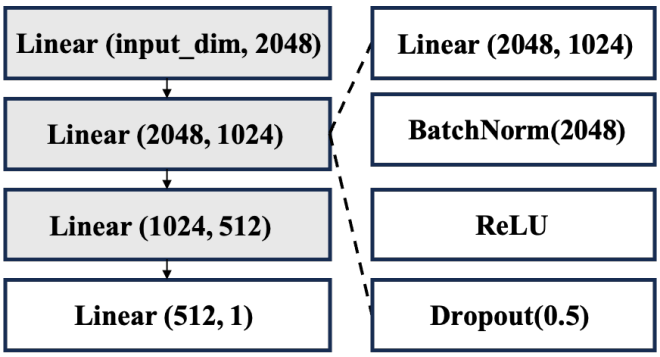


Figure A4. MLP Architecture for Hallucination Detection

To balance the class distribution in the dataset, we randomly sampled from hallucinated samples and clean samples, ensuring that the number of samples in both classes was equal. The final dataset was split into 80% for training and 20% for testing. The training set was used for updating the model parameters, while the test set was used to evaluate the model's performance.

9.3. More Results of Our Method

In this section, we present the results of combining our method with the sampling method, where the temperature parameter is set to 0.7. We simultaneously sample 10 candidate texts and use MLP evaluation to select the text with fewer hallucinations for the next generation. As shown in Tab. A2, compared to the combination with beam search, our method combined with sampling achieves better performance, which fully demonstrates the effectiveness of our method. However, due to the inherent characteristics of the sampling method itself, its recall is slightly lower compared to other methods.

Meanwhile, we conducted experiments on our method using LLaVA-Next-8B. Since the majority of existing methods are evaluated on LLaVA-1.5, we only compared our method with the baseline. The experimental results show that, due to the use of more visual tokens and higher-quality datasets, LLaVA-Next-8B achieves over a 50% reduction in hallucination compared to LLaVA-1.5-7B. Moreover, our experiments on LLaVA-Next-8B also demonstrate that our method remains effective even on a strong baseline.

Method	LLaVA-1.5-7B									
	val2014			val2017			AMBER			
	$C_S\downarrow$	$C_I\downarrow$	$Recall$	$C_S\downarrow$	$C_I\downarrow$	$Recall$	$CHAIR\downarrow$	$Cover\uparrow$	$Hal\downarrow$	$Cog\downarrow$
Beam	54.0	14.8	80.2	47.2	13.0	75.0	9.2	49.9	41.1	4.9
Sample	53.0	15.4	78.5	47.7	13.6	72.8	9.7	51.7	42.9	4.4
(Ours+Beam)	35.2	**9.1**	72.6	37.8	10.5	70.8	6.8	49.4	32.9	3.3
(Ours+Sample)	**32.8**	9.7	67.3	**28.1**	**8.6**	61.3	**6.0**	47.5	**26.7**	**1.7**

Table A2. The experimental results of our method combined with sample search

Method	LLaVA-Next-8B									
	val2014			val2017			AMBER			
	$C_S\downarrow$	$C_I\downarrow$	$Recall$	$C_S\downarrow$	$C_I\downarrow$	$Recall$	$CHAIR\downarrow$	$Cover\uparrow$	$Hal\downarrow$	$Cog\downarrow$
Greedy	26.2	7.3	64.4	26.6	6.6	63.9	5.7	61.8	36.6	2.4
Nucleus	26.6	7.9	63.7	26.0	6.7	63.3	6.0	61.0	37.4	2.9
Beam	26.0	7.5	65.3	24.0	6.0	63.2	5.4	60.7	33.7	2.4
Sample	28.6	8.2	61.7	26.3	7.4	60.6	6.5	57.8	37.3	2.5
(Ours+Beam)	**24.0**	**7.3**	64.0	**22.3**	**5.9**	61.3	**5.2**	57.6	**31.5**	**2.1**

Table A3. The experimental results of our method in LLava-Next-8B.

9.4. Case Study

Additionally, we present a comparison between the text output by our method and that of the baseline. As shown in Fig. A5, the baseline sometimes exhibits hallucinations due to linguistic biases, such as perceiving a 'spoon' when a 'fork' is shown in the image. It also demonstrates inadequate visual understanding, such as mistakenly identifying red leaves as apples. In contrast, our method can mitigate these hallucinations to a certain extent in such scenarios.

10. Datasets and Metrics

CHAIR The CHAIR [37] (Caption Hallucination Assessment with Image Relevance) dataset is a comprehensive evaluation framework for image captioning tasks, which randomly samples 500 images from the MSCOCO2014 validation set. The evaluation process compares the model-generated descriptions against the existing ground truth annotations from the MSCOCO dataset to quantify hallucinations. CHAIR primarily employs two methods to measure hallucination: CHAIR_S at the sentence level, which measures the proportion of hallucinated sentences among all sentences, and CHAIR_I at the instance level, which quantifies the degree of hallucination within a sentence by calculating the ratio of hallucinated objects to all mentioned objects.This evaluation methodology provides a systematic solution for accurately measuring the reliability of image captioning models.

$$\begin{aligned} \text{CHAIR}_\text{I} &= \frac{|\{\text{hallucinated objects}\}|}{\text{all mentioned objects}}, \\ \text{CHAIR}_\text{S} &= \frac{|\{\text{captions with hallucinated objects}\}|}{\text{all captions}} \end{aligned} \tag{9}$$

Unlike previous methods, we follow the publicly available test results of OPERA [29] and use a fixed set of 500 val2014 images for testing instead of random sampling. Additionally, to test the model's performance more comprehensively and definitively, we use the entire set of 4,505 val2017 images for hallucination testing.

AMBER The AMBER [57] is a multidimensional MLLM hallucination evaluation benchmark without relying on LLMs. It collects images from publicly available or copyright-free online databases and refines them through meticulous annotation and filtering. The benchmark includes evaluations for both generative and discriminative tasks, comprising a total of 1,004 images that cover three types of hallucinations: existence, attributes, and relationships.

$$\text{Cover}(R) = \frac{\text{len}(R'_{\text{obj}} \cap A_{\text{obj}})}{A_{\text{obj}}} \tag{10}$$

$$\text{Hal}(R) = \begin{cases} 1 & \text{if CHAIR}(R) \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

$$\text{Cog}(R) = \frac{\text{len}(R'_{\text{obj}} \cap H_{\text{obj}})}{R'_{\text{obj}}} \tag{12}$$

LLaVA-1.5-7B

Beam search:
The image features a calico cat sitting in a pile of leaves and twigs. The cat appears to be staring intently at the camera, capturing the viewer's attention. The cat is positioned in the center of the scene, surrounded by the leaves and twigs.\n\n**There are several apples scattered throughout the scene, with some close to the cat and others further away. These apples add a touch of color and interest to the overall composition of the image.**

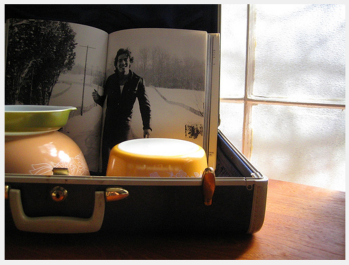
Ours:
In the image, there is a cat sitting in the middle of a pile of leaves. The cat appears to be staring at the camera, capturing the viewer's attention. The leaves are scattered around the cat, creating a cozy and natural environment for the feline to rest in. The cat is positioned in the center of the scene, making it the focal point of the image.

LLaVA-1.5-7B

Beam search:
The image features a close-up view of a slice of cheesecake on a white plate. The cheesecake has a golden-brown crust, and it appears to be a delicious dessert. A fork is placed on the plate, ready to be used to enjoy the cheesecake. \n\n**In addition to the cheesecake, there are two spoons on the plate. One spoon is located on the left side of the cheesecake, while the other spoon is on the right side.**

Ours:
The image features a delicious piece of lemon cheesecake on a plate, ready to be eaten. The cheesecake is sliced, making it easy to serve and enjoy. A fork is placed on the plate next to the cheesecake, indicating that someone is about to dig into this tasty dessert. The cheesecake appears to be well-prepared and visually appealing.

Qwen-VL-7B



Beam search:
The image features an open suitcase on a table, filled with various items. Inside the suitcase, there are two books, one of which is open and has a man on the cover. In addition to the books, there are two bowls placed on top of the suitcase, adding a decorative touch.\n\n **Outside the suitcase, on the table, there is a spoon resting near the open book. The arrangement creates a cozy and unique atmosphere.**

Ours:
The image features an open suitcase on a table, filled with various items such as dishes and a book. The suitcase is in the foreground and takes up the majority of the scene. Inside the suitcase, there are two bowls, one yellow and the other green. These bowls are positioned close to each other, with the green bowl on the left and the yellow one on the right. \n\nOutside the suitcase, a book can be seen resting on top of the table.

Qwen-VL-7B

Beam search:
The image features a wooden bench sitting on a concrete slab in the middle of a desert or empty field. The bench is positioned between a street light and a fire hydrant, making for an unusual setting. **In the distance, there are a few people visible**, but the bench appears to be unoccupied at the moment.\n\nSurrounding the bench, **there are several vehicles parked, including multiple cars and a truck. These vehicles contribute to the desolate atmosphere of the scene.**

Ours:
The image features a deserted area with a bench placed on a concrete slab in the middle of a barren land. The bench is situated close to a street light and faces the direction of the light. There is a fire hydrant located relatively close to the bench as well. \n\nThe surroundings consist of vast, open land with no other visible structures or features. The sky above fills the entire frame of the image, creating an expansive feel to the scene.

Figure A5. Comparison of text generated by our method and the baseline.

Algorithm 1 MLP-Guided Beam Search with Per-Candidate Window Tracking

Input: Multimodal context $\mathbf{x}_{\text{mm}}$, MLLM F, max length $L_{\max}$, beam width B, `[EOS]`, `[BOS]`, window size W, MLP scorer $\mathcal{M} : \mathbb{R}^{d\times W} \to \mathbb{R}$
Output: Generated sequence $\mathbf{y}^*$

Initialize beams: $\mathcal{B} \leftarrow \{(\mathbf{y} = [\texttt{[BOS]}],\ s = 0.0,\ \mathcal{A} = [],\ w = 0)\}$
$\mathcal{F} \leftarrow \emptyset$ ▷ Finished sequences
for $t = 1$ **to** $L_{\max} - 1$ **do**
  $\mathcal{C} \leftarrow \emptyset$
  for each $(\mathbf{y}, s, \mathcal{A}, w) \in \mathcal{B}$ **do**
    if $\mathbf{y}[-1] = \texttt{[EOS]}$ **or** $w = W$ **then**
      $\mathcal{F} \leftarrow \mathcal{F} \cup \{(\mathbf{y}, s, \mathcal{A}, w)\}$
      continue
    end if
    $\mathbf{x}_{\text{in}} = [\mathbf{x}_{\text{mm}}; \mathbf{y}]$
    $(\mathbf{z}, \mathbf{att}) = F(\mathbf{x}_{\text{in}}, \text{return\_attentions=True})$
    $\mathbf{z}_{\text{next}} = \mathbf{z}[-1], \mathbf{a}_{\text{next}} = \mathbf{att}[-1]$
    $\mathbf{p} = \log \text{softmax}(\mathbf{z}_{\text{next}})$
    Get top-B tokens: $\{(y_k, \ell_k)\}_{k=1}^{B} = \arg\text{top}_B(\mathbf{p})$
    for $k = 1$ **to** B **do**
      $\mathbf{y}' \leftarrow \mathbf{y} \circ [y_k]$
      $s' \leftarrow s + \ell_k$
      $\mathcal{A}' \leftarrow \mathcal{A} \circ [\mathbf{a}_{\text{next}}]$
      $w' \leftarrow w + 1$
      $\mathcal{C} \leftarrow \mathcal{C} \cup \{(\mathbf{y}', s', \mathcal{A}', w')\}$
    end for
  end for
  if $\mathcal{C} = \emptyset$ **then**
    break
  end if
  $\mathcal{B} \leftarrow \text{top}_B(\mathcal{C}, \text{key} = s')$
    ▷ Check if all active beams have completed a window
  if all $(\mathbf{y}, s, \mathcal{A}, w) \in \mathcal{B}$ satisfy $w = W$ **then**
      ▷ MLP rescoring on last W attentions
    for each $(\mathbf{y}, s, \mathcal{A}, w) \in \mathcal{B}$ **do**
      $\mathbf{A}_{\text{win}} = [\mathcal{A}[-W], \ldots, \mathcal{A}[-1]] \in \mathbb{R}^{d\times W}$
      $s_{\text{mlp}} = \mathcal{M}(\mathbf{A}_{\text{win}})$
      Associate s_{mlp} with candidate
    end for
    $(\mathbf{y}^{\text{best}}, s^{\text{best}}, \mathcal{A}^{\text{best}}, w^{\text{best}}) \leftarrow \arg\max s_{\text{mlp}}$
    $\mathcal{B} \leftarrow \{(\mathbf{y}^{\text{best}}, s^{\text{best}}, \mathcal{A}^{\text{best}}, 0)\}$ ▷ Reset window counter
    $\mathcal{F} \leftarrow \emptyset$ ▷ Clear finished, continue from best
  end if
  if all $(\mathbf{y}, s, \mathcal{A}, w) \in \mathcal{B}$ satisfy $\mathbf{y}[-1] = \texttt{[EOS]}$ **then**
    break
  end if
end for
$\mathcal{F} \leftarrow \mathcal{F} \cup \mathcal{B}$
return $\mathbf{y}^* = \arg\max_{(\mathbf{y}, s, \mathcal{A}, w)\in\mathcal{F}} s$

The dataset has four evaluation metrics. CHAIR is used to measure the frequency of the appearance of hallucinatory objects in the responses of generation tasks. Cover quantifies the proportion of object coverage in the responses. Hal represents the proportion of responses with hallucinations. Cog evaluates the possibility of generating specific hallucinatory objects. These four metrics assess the hallucination situations of multimodal large language models in generation tasks from different dimensions.

POPE POPE [35] was also constructed based on the MSCOCO for evaluating object hallucination in large vision-language models, which provides a robust foundation for hallucination assessment. The evaluation dataset, through a combination of automatic and manual annotations, randomly selects images containing more than three objects to pose simple Yes-or-No questions to the model. To evaluate object hallucination in Large Vision-Language Models (LVLMs), researchers designed three representative sampling strategies: (1) Random Sampling strategy serving as an unbiased baseline method; (2) Popular Sampling strategy, which focuses on objects with high occurrence frequency in the dataset; and (3) Co-occurrence Sampling strategy, which considers contextual relationships between objects. Each type of data is based on the same set of 500 images, with six questions posed for each image.

GPT-4o evaluation: Following the experimental setup by Huang et al. [29], we use GPT-4o to evaluate the accuracy and detailedness of the answers, providing a more nuanced assessment of the generated responses. For two different model architectures, we compare our best method with others. Notably, we do not perform random sampling and instead evaluate all 500 images.

11. Limitation

Finally, we intend to delve into the limitations of this study. To begin with, our method centers on object hallucinations, and more robust solutions remain to be developed to tackle positional relationship hallucinations. Additionally, employing GPT-4o for hallucination evaluation constitutes only a suboptimal approach, and reliable methodologies need to be sought in the future to measure hallucinations in MLLMs.