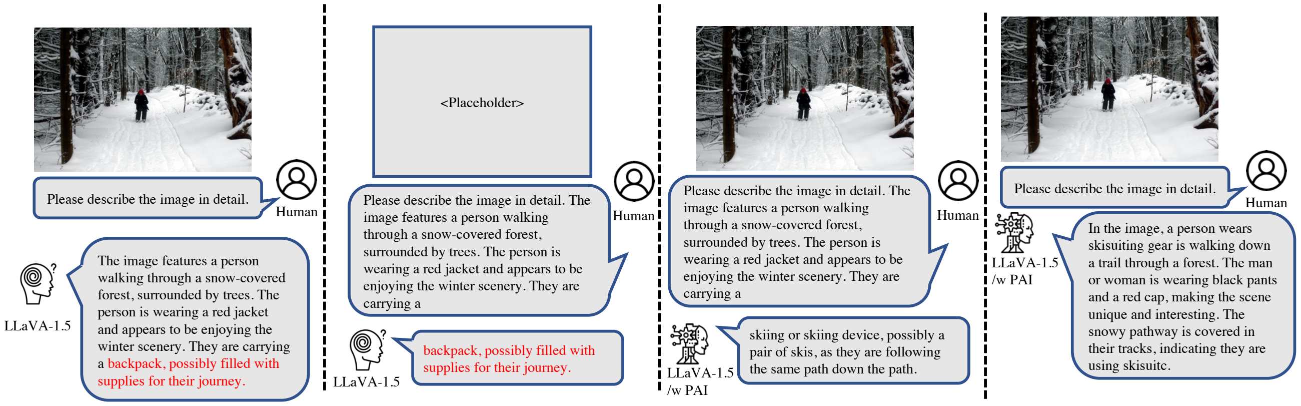
Paying More Attention to Image: A Training-Free Method for Alleviating Hallucination in LVLMs

Shi Liu Kecheng Zheng Wei Chen

State Key Lab of CAD&CG, Zhejiang University



Please describe the image in detail.

The image features a person walking through a snow-covered forest, surrounded by trees. The person is wearing a red jacket and appears to be enjoying the winter scenery. They are carrying a backpack, possibly filled with supplies for their journey.

Please describe the image in detail. The image features a person walking through a snow-covered forest, surrounded by trees. The person is wearing a red jacket and appears to be enjoying the winter scenery. They are carrying a

backpack, possibly filled with supplies for their journey.

Please describe the image in detail. The image features a person walking through a snow-covered forest, surrounded by trees. The person is wearing a red jacket and appears to be enjoying the winter scenery. They are carrying a

skiing or skiing device, possibly a pair of skis, as they are following the same path down the path.

Please describe the image in detail.

In the image, a person wears skisuiting gear is walking down a trail through a forest. The man or woman is wearing black pants and a red cap, making the scene unique and interesting. The snowy pathway is covered in their tracks, indicating they are using skisuitc.

(a) Image description by LLaVA **(b)** Same hallucinations without image **(c)** Correct description with same history **(d)** Image description by PAI

Figure 1. We present an examination of various input settings, with hallucinations specifically highlighted in red. (a) When using LLAVA for image description, it generates a hallucinated description. (b) Even without image input, when only the historical response preceding the hallucinated description is input to LLAVA, it reproduces the same hallucinated description, a phenomenon we refer to as "Text Inertia". (c) Our proposed method, PAI, effectively mitigates this text inertia problem and yields accurate descriptions. (d) Utilizing PAI for image description results in a significantly more precise description.

Abstract

Large Vision-Language Models (LVLMs) align image features to the input of Large Language Models (LLMs), enhancing multi-modal reasoning and knowledge utilization capabilities. However, the disparity in scale between models of different modalities has resulted in LLMs assuming a predominant role in multimodal comprehension. This imbalance in model integration can lead to instances of hallucinatory outputs. In particular, LVLMs may generate descriptions that persist in the absence of visual input, suggesting that these narratives are disproportionately influenced by the textual context. We refer to this phenomenon as "text inertia." To counteract this issue, we introduce a training-free algorithm designed to find an equilibrium between image comprehension and language inference. Specifically, we firstly involve adjusting and amplifying the attention weights assigned to image tokens, thereby granting greater prominence to visual elements. Meanwhile, we subtract the logits of multimodal inputs from the model logits of pure text input, which can let model not biased towards only LLM. By enhancing images tokens and reducing the stubborn output of LLM, we can let LVLM pay more attention to images, towards alleviating text inertia and reducing the hallucination in LVLMs. Our extensive experiments shows that this method substantially reduces the frequency of hallucinatory outputs in various LVLMs in terms of different metrics[1].

1. Introduction

Recently, Large Vision-Language Models (LVLMs) have made significant strides, exhibiting impressive capabilities across a multitude of tasks [13, 17, 38, 44]. How-

---

[1] The source code is publicly available at: https://github.com/LALBJ/PAI

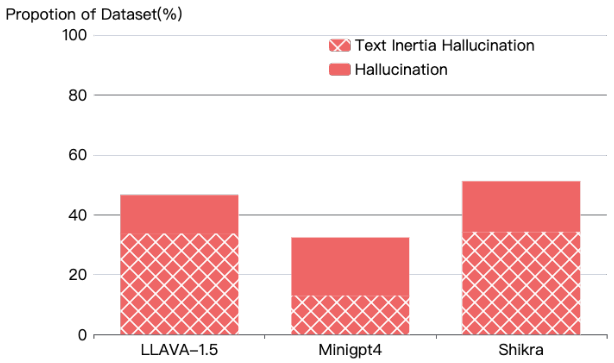


Figure 2. Percentage of hallucination and percentage of text inertia hallucination(calculated with 500 samples). For specific calculation processes, please refer to the supplementary material.

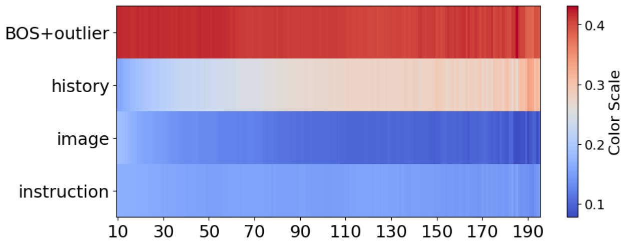


Figure 3. Visualization of the average attention ratio of different content. The x-axis denotes the sequence length of the history tokens. The lengths of the image, instruction, BOS, and outlier tokens [11] are all fixed as they are part of the model input, with 576 for the image, 21 for the instruction, and 1 each for the BOS and outlier tokens.

ever, these models still struggle with the phenomenon of hallucination, where there is a mismatch between the textual content generated by the model and the actual visual input it receives [27].

Hallucination in LVLMs is often attributed to issues with modality alignment, leading to the development of mitigating strategies through alignment training optimization [9, 22, 34]. However, is hallucination in LVLMs merely a result of the model's capacity, and can it only be alleviated through additional training? We propose a scenario where LVLMs generate a hallucinated object description. Specifically, even when the image input is removed and only the generated text preceding the hallucinated object word is retained, the LVLMs persist in producing the same hallucinated description, as depicted in Fig. 1.

In order to empirically investigate this behavior, we conducted tests on three LVLMs within the context of image describing tasks on the COCO dataset. We identified and conducted a statistical analysis on instances where LVLMs generated identical hallucinated object descriptions, even when the input was exclusively historical response text without any image. The observation from Fig. 2 clearly indicates that, even with the application of rigorous identification settings, the phenomenon continues to represent a substantial proportion.

We referred this phenomenon as **"Text Inertia"**. Our hypothesis is that text inertia arises due to the current generative paradigms mapping image representations onto the text representation space as text tokens. In this mechanism, the LLM becomes the dominant character, and the inference process lacks additional handling of image tokens, leading to their neglect during the generation process. To validate this hypothesis, we have analyzed the attention values ratios of the LLaVA model during the inference process in Fig. 3. Our findings show that despite image tokens occupying a significant proportion, they do not receive substantial attention under the current mechanism. This multimodal chat resembles more of an automatic completion based on context rather than a continuous attention to the image for completion.

To close this gap, we introduce a method refered to as **Pay Attention to Image (PAI)**. At a high level, PAI intervenes in the inference process to make it more image-centric, following the original image perception direction. To achieve this, we focus on the self-attention heads in the decoder layers of LVLMs. We enhance the activation values for image tokens in their original directions during inference. This allows us to use the updated attention matrix to calculate the hidden states for the generated token, thereby incorporating more consideration for image representation during the generation process. To further mitigate text inertia, we construct the input using instruction text and historical response tokens, and subtract the model logits of this input from the logits of our intervened model. This strategy helps to reduce the influence of language priors during the generation process. Unlike previous methods for mitigating hallucination that require additional training or external tools, our approach is training-free. Moreover, we are the first to propose an inference intervention method for mitigating hallucination in LVLMs.

We then evaluate the response accuracy in the image description task through the CHAIR metric [31] and gpt-4v. In addition, we use POPE [21] and MMHal-Bench [33] to more comprehensively evaluate the model's hallucination performance on VQA task. Since our model intervenes in the inference process, it can be used for any decoding method. Therefore, we conducted experiments on three decoding methods of the three models. The experimental results proved the effectiveness of our method in mitigating hallucinations.

In summary, the contributions can be summarized as follows:

1. We demonstrate the phenomenon of text inertia that LVLMs answer only by language context without looking at the image.

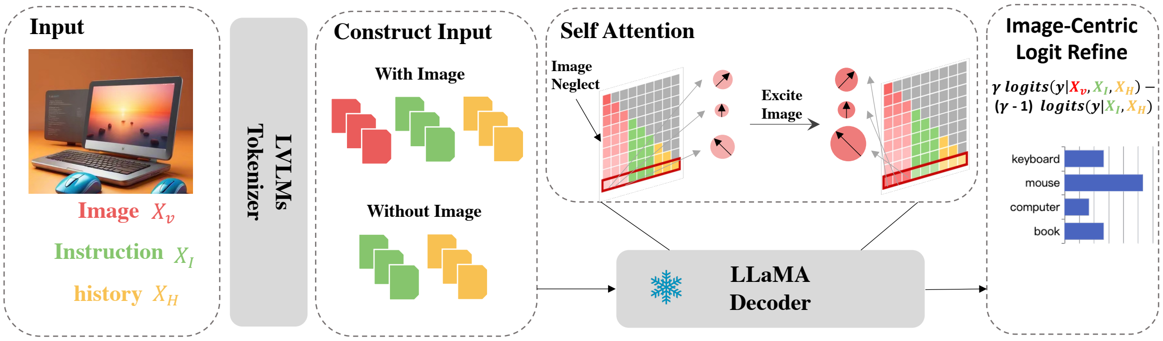


Figure 4. The framework of our method. It demonstrates the inference process of a single token within our approach. To alleviate text inertia, we additionally construct an input without image. Throughout the forward inference process, we amplify the focus on the image token by edit self-attention maps in LLaMA. Ultimately, we subtract the logits distribution of the language prior during decoding to achieve an accurate description.

2. Based on the above analysis, we further found that the hallucination behavior is related to the neglect of images. Therefore, we further proposed a inference intervention algorithm to make LVLMs pay more attention to the image.

3. Through extensive experiments, we have demonstrated the effectiveness of PAI in mitigating hallucination problems in a training-free manner.

2. Related Work

2.1. Large Vision-Language Models

The development of pre-training techniques [5, 29] and instruction tuning techniques [28, 39] has rapidly advanced LLMs technology, such as LLaMA [35] and Vicuna [30], further leading to the prosperity of LVLMs technology. Early works, like Flamingo [2] and BLIP-2 [19], have successfully adapted LLMs to visual tasks, demonstrating notable generative capabilities and in-context learning abilities. Recently, the capabilities of LVLMs have further advanced under the influence of visual instruction tuning techniques [25, 26]. Using different projectors to map images to the text domain, thereby endowing language generation models with image understanding capabilities, is also a hot research topic [10, 26, 41, 46]. Additionally, several studies focus on visual language tasks such as grounding capabilities [8] and reasoning capabilities [17]. However, recent LVLMs still face the issue of hallucination generation [23].

2.2. Mitigation of LVLMs Hallucination

Hallucination in LVLMs refers to contradictions between the image input and the textual output. Various methods have been proposed to mitigate hallucination. The most direct reason for hallucination generation is that hallucination arises from data bias and knowledge gaps between vision and language. Therefore, better data filtering methods [14, 24, 43] and higher quality annotated data [3] are introduced. Simultaneously, these methods also imply the need for more alignment training [34] or adjustments in the model architecture [9, 22]. These methods can achieve good results, but they are time-consuming and require high computational resources.

Apart from addressing the ability of LVLMs itself, hallucination can also be mitigated through post-processing methods. This method usually involves using an additional module or external tools to edit the response. Recent methods such as LURE [45] utilize additional data to train a state detector and when hallucination issues are detected, content is regenerated by a revisor model. Woodpecker [42] introduces an external visual model to inspect entities extracted from the response, and then the detection results are handed over to the generation model to regenerate better answers. These methods also extend the inference chain and increase inference costs.

Training-free hallucination mitigation methods have so far only been attempted in decoding methods. OPERA [15] discovered an abnormal attention pattern that accompanies model decoding. It was statistically found that this pattern often accompanies hallucination descriptions, and thus a detection and mitigation method was proposed based on this pattern to alleviate the hallucination faced by the model. VCD [18] introduced the notion that visual uncertainty increases hallucination descriptions and, based on this discovery, proposed a contrast decoding method to alleviate hallucination issues.

3. Preliminaries

The architecture of LVLMs typically comprises of three main components: an image encoder, a projector, and a language decoder. Both the image encoder and language decoder are usually pre-trained. The image encoder is employed to transform images into image tokens, which are subsequently mapped to the text representation space by the projector. This process enables the concatenation of image tokens with text tokens that are then fed into the language decoder. The language decoder subsequently generates corresponding responses based on the provided instructions.

The existing projectors. Currently, projectors predominantly fall into two categories: linear projectors and resamplers. A projector takes N visual features from the image encoder and transforms them into M visual tokens. The linear projector employs a multilayer perceptron to transform all visual features, maintaining a one-to-one transformation which meaning that M equals N. In contrast, the resampler does not preserve all visual features but instead samples visual cues (M, where M < N). For instance, Q-former [19] utilizes M learnable queries and Bert [12] to extract information from visual features. Given that the knowledge of images during the generation process solely originates from the output image tokens of the projectors, our attention is concentrated on the image tokens post-projection, irrespective of their preceding modeling process.

Autoregressive language decoders. Nearly all LVLMs adopt LLaMA-family models as their language decoders, which employ the self-attention mechanism. The visual tokens processed by the projector are concatenated with text tokens and fed into the LLaMA, which carries out the forward decoding process. From a single-layer perspective, each layer repetitively performs the following operation with the same input shape:

$$\text{head}_h = \text{softmax}\left(\frac{Q_h K_h^T}{\sqrt{d_k}}\right) V_h, \tag{1}$$

$$X_{l+1} = X_l + \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_H) W_O. \tag{2}$$

In these equations, X_l represents the hidden states output from the l-th layer of the network, reflecting the intermediate state of the information being processed by the model at layer l. Each head_h performs an attention operation using its own set of queries Q_h, keys K_h, and values V_h. This operation enables the model to focus on different parts of the input for each head, thereby capturing various aspects of the knowledge in the input. The final output is $y \in R^{1\times V}$, where V is the length of the vocabulary, representing the probability of each token in the vocabulary. In each iteration, a token is generated. The sequence generation continues until an EOS token is produced, marking the end of generation and resulting in a complete response.

4. Method

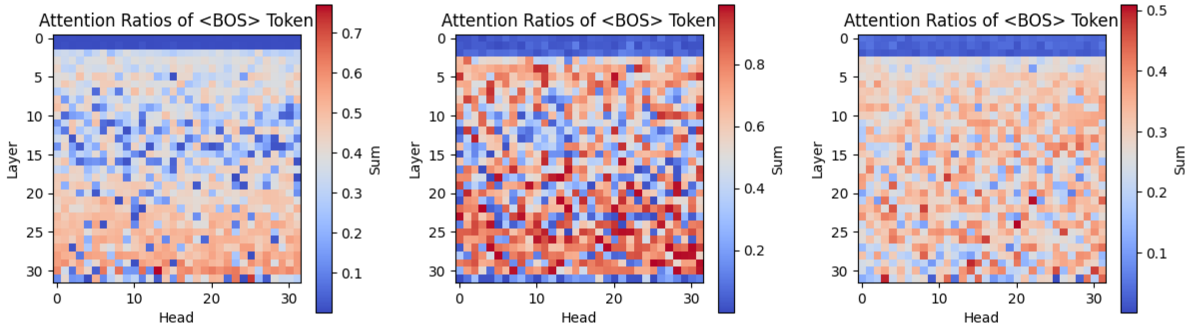
At the core of our method is a solution for image neglect and text inertia, both of which are fundamentally interconnected. Essentially, as paying more attention to the image, there is a corresponding reduction in the reliance on language priors. Intuitively, in a conversation centered around an image, the model should devote more attention to the image, thereby allowing it to have a significant impact on the response. As such, we identify the self-attention map in the token-level generation and augment the image attention in its original directions. This strategy promotes a more image-centric latent representations. Additionally, to further mitigate the influence of text inertia, we devide the logits distribution of pure text input into the model's output.

4.1. Pay More Attention to Image

Extracting the Self-Attention Matrix. We start from a token-level perspective. The response process in LVLMs is fundamentally generated token by token. Each token is generated based on the input image, instruction, and the historically generated response. This process is facilitated through a multi-layer attention decoder architecture. Consequently, this results in a probability distribution of the vocabulary for the currently generated token. Our goal is to extract the attention matrix of each attention head at every layer, indicating the influence of each content during inference.

When generating the k-th token in the sequence, the input representation for the attention head in the forward process includes the instruction representation $X_I = \{x_{i_1}, ..., x_{i_n}\}$, image representation $X_V = \{x_{v_1}, ..., x_{v_n}\}$, and the representation of the historically generated response $X_H = \{x_{h_1}, ..., x_{h_k}\}$. Notably, the image representation considered here is the one that has been processed by the projector. Essentially, the hidden states of each input layer are $X = \text{concat}(X_I[1 : m], X_V, X_I[m + 1 : n], X_H)$, where the notation $X_I[1 : m]$ indicates the first m elements in the instruction representation. Each attention head assigns different degrees of attention to each element during the current token representation encoding process. Our aim is to enhance the attention paid to the image. Therefore, we extract the attention matrix values $A_{h_k, x_{v_1}:x_{v_n}}$ related to the image in the current generated token process, intervene, and then redistribute the attention values of each element through softmax.

Excite model in a trustful direction. There have been attempts in some LLMs works to make the answers generated by LLMs more trustworthy by means of intervention [4, 6, 16, 20, 36]. The implemented approach typically involves intervening with the hidden states. As for defining what constitutes a more trustworthy direction, it usually requires additional projection and training to probe this trustful direction. In our case, a response that is more



(a) LLAVA (b) Minigpt4 (c) Shikra

Figure 5. The BOS token attention ratios of three model.

image-based is considered more trustworthy. Since LVLMs have undergone alignment training, the original attention values provide a direction based on image content. As illustrated in Fig. 4, by amplifying the attention values of image tokens based on the original attention values, we can enhance the trustworthiness of our results.

Another nuance involves our avoidance of choosing the attention head used to shift activation. In the ITI method [20], it is stated that not all attention heads should be subjected to intervention. Therefore, they introduce a trustful score to rank each head across all layers and select the top-k heads for intervention. In our case, the less trustworthy heads with lower attention values receive less intervention. We first extract the image tokens' attention values $A \in$ for the generated token. We then use the hyper-parameter α to control the step size for intervention. From a single attention head perspective, our method can be expressed as follows:

$$A_{h_k,x_{v_1}:x_{v_n}} = A_{h_k,x_{v_1}:x_{v_n}} + \alpha * |A_{h_k,x_{v_1}:x_{v_n}}|. \quad (3)$$

The model's final vocabulary probability distribution is derived from the projection of the hidden states of the last token in the sequence. Therefore, we extract the attention values of the last token regarding the image token by indexing h_k. Following the intervention, we use the $softmax$ function to redistribute the attention values of each token during the reassignment of encoded hidden states. This procedure is repeated for each subsequent token prediction in an autoregressive manner and is independent of the choice of the decoding algorithm.

Excite more precisely with an attention mode prior. The presence of the BOS token, an attention sink pattern [40], in a sentence results in higher attention values during the attention computation process, which may seem counterintuitive. The BOS token typically signifies the start of a sentence and, as such, does not carry significant semantic content. However, the generation of tokens is significantly influenced by this particular token, a similar pattern that also manifests itself in visual models [11]. As mentioned in StreamLLM [40], the pattern of the attention sink emerges when redundant attention values are present.

Naturally, one might infer that when the sink pattern appears, we excite the image token. To further investigate this phenomenon, as depicted in Fig. 5, we find that the sink phenomenon is not overtly evident in the shallow layers. This is because the shallow layers tend to focus more on encoding semantically rich information [37]. When the encoding of semantically rich tokens stabilizes, the attention sink phenomenon arises. Therefore, we build upon the judgement of intervention timing by calculating the similarity of the hidden states.

4.2. Image-Centric Logit Refine

In Fig. 1, we observe a peculiar phenomenon where LVLMs, even after completing their hallucination descriptions, continue to generate identical hallucinated text even when the image is removed from the input. This observation naturally leads us to the concept of using the output distribution (when no image is in input) as a reference to penalize our initial prediction distribution. The final output of our LVLMs is the predicted probability for each token.

$$p(y|X_V, X_I, X_H) = \gamma p(y|X_V, X_I, X_H) - (\gamma - 1) p(y|X_I, X_H). \quad (4)$$

This equation effectively reduces the predicted probability based on text alone. The weight γ is used to control the degree of penalty applied to the initial prediction distribution.

This operation is conceptually similar to LLM-CFG [32]. Essentially, it provides a guided generation mechanism that allows the model to make informed choices between outputs based on image content and those based on language logic. This way, the model can better balance the influence of visual and textual information in its outputs, leading to more contextually accurate and relevant results.

5. Experiments

5.1. Setup

Baselines. We evaluate the effectiveness of our method on three different models. To better compare the impact of image feature tokens after different projectors on our method, we selected two models that use linear projectors, LLAVA and Shikra, as well as one model that uses resamplers, Minigpt4. Additionally, for a more convincing comparison, we report on three decoding methods for comparison: greedy, beam search, and nucleus sample. We also selected the OPERA [15] method, which is an improvement on beam search, and the VCD [18] method, which is an improvement on nucleus sampling, to compare with our results. We used the default hyperparameters from the open-source versions of these two methods.

Implementation Details. As different models have different lengths of image tokens, leading to different degrees of image neglect, to better align with the image sequence length of the model, we set $\alpha = 0.5$ for LLAVA, $\alpha = 0.6$ for Shikra with long image token sequence lengths and $\alpha = 0.2$ for resampler models with short image token sequences. As text inertia is independent of the image token length, we continuously use $\gamma = 1.1$. Apart from this, in the beam search tests, the beam number is set to 5 for all methods, and in the nucleus sample tests, all the common parameters are consistent.

5.2. Benchmark & Evaluation Metrics

CHAIR. Caption Hallucination Assessment with Image Relevance (CHAIR) is a widely used metric in image captioning tasks. CHAIR operates by creating a set of ground-truth object labels for each image. Any object mentioned in the caption that does not exist in the label set is considered a hallucinated object. CHAIR comprises two evaluation dimensions: instance-level and sentence-level, represented as $\mathrm{CHAIR_I}$ and $\mathrm{CHAIR_S}$, respectively. These are calculated in the following manner:

$$\mathrm{CHAIR_I} = \frac{|\{\text{hallucinated objects}\}|}{\text{all mentioned objects}}, \tag{5}$$

$$\mathrm{CHAIR_S} = \frac{|\{\text{captions with hallucinated objects}\}|}{\text{all captions}}. \tag{6}$$

$\mathrm{CHAIR_I}$ represents the proportion of hallucinated objects in all objects within a sentence, while $\mathrm{CHAIR_S}$ refers to the proportion of hallucinated sentences, i.e., image descriptions containing hallucinations, in all images. We conducted experiments based on the validation set of MSCOCO 2014 with randomly sampled 500 instances. By prompting the LVLMs with "`Please help me describe the image in detail.`", we subsequently employ the CHAIR metric to evaluate the generated description. Meanwhile, considering the impact of sequence length on the hallucination, we set max_new_tokens to 512 for all models and decoding method.

POPE. The Polling-based Object Probing Evaluation (POPE) is a evaluation metric designed in the VQA paradigm. POPE serves as a metric for assessing object hallucination, evaluating hallucinations by asking LVLMs questions such as "`Is there a <object> in the image?`" Here, `<object>` is replaced with the constructed ground-truth object and three different types of split objects. In the "random" split, objects are randomly selected from the entire dataset for evaluation. In the "popular" split, objects are chosen from those most frequently appearing in the dataset. In the "adversarial" split, objects that are highly related to the image objects are selected for evaluation. We conduct our evaluation on the COCO dataset with 500 images, with each image having 6 questions for each type of POPE. We evaluate the performance of the model in object recognition tasks using both accuracy score and F1 score.

MMHal-Bench. For further evaluation of our method on some challenging datasets, we choose MMHal-Bench, which is designed with 96 image-question pairs, spread across 8 question categories × 12 object topics. It contains eight types of questions about object attributes, adversarial objects, comparisons, counting, spatial relations, environment, holistic descriptions, and others to comprehensively assess the model's hallucination performance on high-difficulty datasets. Essentially, it is also a VQA-based evaluation, but unlike the existence-based examination in POPE, its questions also include some logical considerations. Therefore, for evaluation on MMHal-Bench, we first need to answer questions and then use GPT-4 to score the answer based on the response and the ground-truth answer. The evaluation results include the model's scores across all question categories, and the overall score represents the average of these scores.

GPT-4v Assisted Evaluation. To further evaluate the model's performance in image description tasks, we can move beyond the CHAIR metric, which is based on information extraction and only considers object hallucination. We can use GPT-4v for open evaluation. As with previous evaluations [15, 18], we sample 50 images on the COCO dataset for evaluation. We construct prompts, input images into GPT-4v, along with the descriptions responses from two assistants. GPT-4v evaluation takes into account two dimensions: Accuracy and Detailedness, denoted respectively as C and D. Detailed prompt construction can be found in the appendix.

5.3. Experimental Results

In this section, we analyze the performance of PAI across various hallucination evaluation tasks, including long image description, simplified VQA answer, construction of metric

Table 1. CHAIR hallucination evaluation results on three models. Smaller values corresponds to less hallucinations

Decoding	**Method**	**LLAVA**		**Minigpt4**		**Shikra**	
		$\mathrm{CHAIR}_\mathrm{S}$	$\mathrm{CHAIR}_\mathrm{I}$	$\mathrm{CHAIR}_\mathrm{S}$	$\mathrm{CHAIR}_\mathrm{I}$	$\mathrm{CHAIR}_\mathrm{S}$	$\mathrm{CHAIR}_\mathrm{I}$
Greedy	Vanilla	46.6	13.4	32.8	11.1	51.2	14.4
	PAI	**24.8**	**6.9**	**26.3**	**8.8**	**37.6**	**10.0**
Beam Search	Vanilla	46.4	14.3	46.6	13.4	53.0	14.7
	OPERA	44.6	14.4	30.1	9.8	36.8	12.4
	PAI	**21.8**	**5.6**	**24.8**	**6.9**	**35.8**	**11.4**
Nucleus	Vanilla	58.2	18.2	32.7	11.9	57.9	16.4
	VCD	51.8	15.1	34.8	11.5	57.6	16.3
	PAI	**43.4**	**14.7**	**26.7**	**10.3**	**49.9**	**13.2**

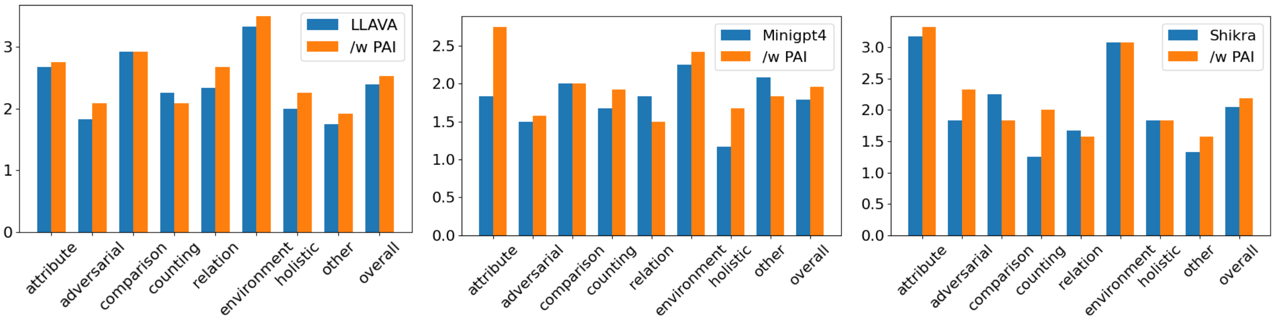
Table 2. Results on POPE. The best performances within each setting are **bolded**.

Model	**Decoding**	**Method**	**Random**		**Popular**		**Adversarial**	
			Acc	F1	Acc	F1	Acc	F1
LLAVA	Greedy	Vanilla	89.38	89.61	85.9	86.3	79.0	80.88
		PAI	**90.3**	**90.4**	**87.03**	**86.9**	**81.06**	**81.96**
	Beam Search	Vanilla	88.10	87.87	85.5	85.21	81.2	81.69
		PAI	**88.72**	**88.24**	**86.57**	**85.81**	**82.2**	**82.04**
	Nucleus	Vanilla	82.9	80.87	81.13	79.26	78.67	77.16
		VCD	83.17	84.04	**83.63**	82.31	**81.1**	**80.12**
		PAI	**84.0**	**84.67**	83.6	**84.35**	80.6	79.6
Minigpt4	Greedy	Vanilla	79.6	78.39	72.8	72.13	70.78	70.82
		PAI	**81.0**	**81.11**	**73.6**	**74.71**	**71.0**	**72.89**
	Beam Search	Vanilla	76.35	72.23	72.22	68.26	70.2	66.74
		PAI	**77.25**	**74.06**	**72.6**	**69.69**	**70.4**	**68.03**
	Nucleus	Vanilla	61.23	64.75	57.26	61.54	55.8	60.74
		VCD	64.05	63.34	59.77	60.23	59.33	60.93
		PAI	**64.36**	**67.37**	**60.5**	**64.38**	**59.64**	**63.75**
Shikra	Greedy	Vanilla	83.57	**83.91**	83.1	83.18	79.1	**79.99**
		PAI	**84.41**	83.65	**84.2**	**83.7**	**80.33**	79.49
	Beam Search	Vanilla	85.19	84.99	82.8	82.55	80.03	80.3
		PAI	**85.49**	**85.2**	**83.7**	**83.1**	**80.53**	**80.6**
	Nucleus	Vanilla	82.4	82.87	81.6	**81.91**	76.7	77.79
		VCD	82.13	82.37	79.83	79.95	77.53	78.28
		PAI	**83.6**	**83.78**	**82.2**	81.75	**77.43**	**78.7**

evaluation, and leveraging the near-human cognitive capabilities of GPT-4/GPT-4v as evaluation methods. For further analysis, please refer to the appendix.

Results on CHAIR. The experimental results are presented in Tab. 1. As our approach is an inference intervention method, it differs from previous decoding hallucination mitigation methods that primarily concentrate on improving a single decoding method. We have tested our method on three decoding techniques. Our method has achieved hallucination mitigation on all three decoding methods used by the three models. However, when integrated with the nucleus, a sampling-based method, the hallucination reduction brought about by our method is not significant. This may be because even though our method has increased the priority of trustful tokens, the sample set during nucleus decoding still contains many hallucination tokens.

Moreover, while OPERA significantly mitigates hallucinations, its time efficiency is considerably higher compared



(a) LLAVA (b) Minigpt4 (c) Shikra

Figure 6. The performance of the three models on the MMHal-Bench. Higher scores indicate superior performance.

Table 3. Results on GPT-4V evaluation.

Method	**LLAVA**		**Minigpt4**		**Shikra**	
	C	D	C	D	C	D
Greedy	5.62	5.24	5.8	5.74	5.54	**5.25**
PAI	**6.46**	**5.36**	**7.04**	**5.89**	**6.04**	5.05

to vanilla. In contrast, our method not only has almost the same time efficiency as vanilla, but it also performs better in reducing hallucination issues. Compared to VCD, during the generation process of long sequence tasks, the introduction of visual uncertainty during decoding sometimes leads to more hallucination descriptions. However, our method can reduce the proportion of hallucination words in the sample pool.

Results on POPE. Unlike the CHAIR evaluation, POPE is in a VQA format, so the response is brief, answering only “Yes” or “No”. The phenomena of text inertia and image neglect may not be as noticeable under this setting. However, our method still achieved a notable improvement compared to the vanilla decoding method. Moreover, compared to methods like VCD that focus on VQA hallucination with visual uncertainty, our approach remains competitive. As for OPERA, due to time efficiency issues and the explain in its paper that knowledge aggregation patterns are negligible in short sequence tasks, the effect is minimal. Therefore, we have not included the OPERA method in our comparison.

Results on MMHal-Bench. The experimental results, as shown in Fig. 6, indicate that for some more image-based question types, such as object attributes, adversarial objects, and holistic questions, the answers are more accurate when inference intervention with PAI is applied, and there is a certain degree of improvement across all models. However, for some logical questions, such as comparisons and relations, there is no noticeable improvement after intervention. In summary, through the overall metric, i.e., the average of the eight evaluation dimensions, there is a certain degree of improvement compared to the baseline after incorporating PAI.

Results on GPT-4v Assisted Evaluation. The experimental results, as shown in Tab. 3, indicate that even when more comprehensive hallucination evaluation dimensions are added, our method, compared to the greedy decoding method, can provide more accurate responses on all three models without losing detail in the description. Given that GPT-4v’s visual understanding and language logic capabilities have reached a level close to that of humans, it can more comprehensively illustrate the performance improvements brought about by our method.

Results on qualitative cases. Fig. 7 presents two cases for comparison. In the first kitchen scene, LLAVA provides accurate object descriptions initially, but hallucinates the positions. However, as the history information expands, LLAVA tends to describe objects related to the kitchen scene, such as bowls and dishwashers, driven by textual inertia. Our proposed PAI, by mitigating this textual inertia and increasing focus on the image, generates more accurate results. Additionally, the second case demonstrates the enhancement in model’s counting ability achieved by our method.

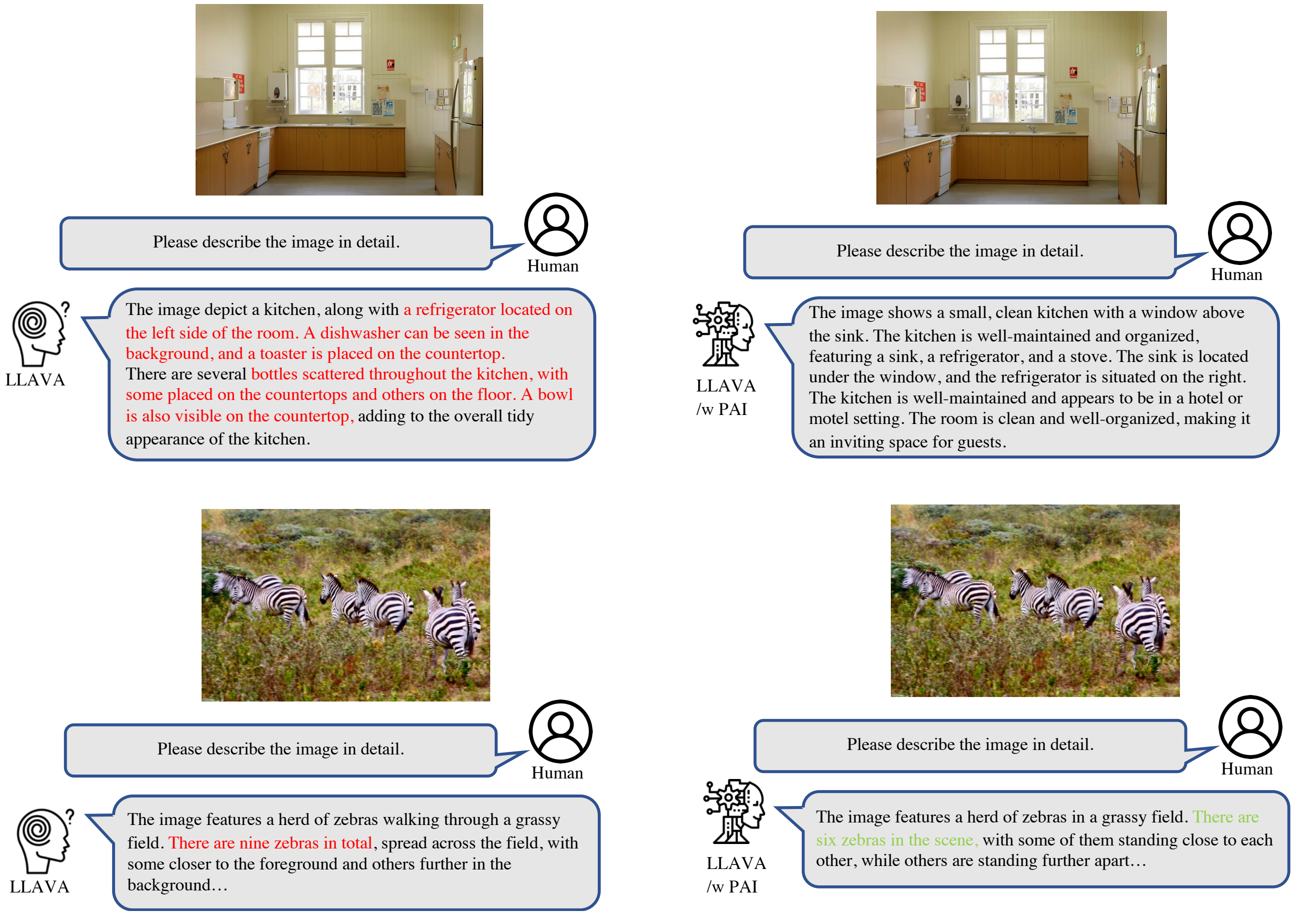
5.4. Ablation Study

Our method, PAI, consists of two stages of interventions. In the first stage, during forward inference, the hyperparameter α is utilized to set the scale of intervention. Simultaneously, the layer prior, represented as “L”, is used to determine the attention layer for intervention. The second stage unfolds during the decoding process, where we mitigate text inertia by subtracting the logits distribution that results from inputs devoid of image information. In this stage, the scale is managed by the parameter γ.

We use LLaVA-1.5 as the representative LVLM baseline and the greedy decoding method as the basic baseline to compare the impact of our hyperparameters on the task of long sequence image description. To evaluate our method, we choose the CHAIR metric. However, since CHAIR only evaluates the hallucination problem, we have incorporated the F1 score to consider information richness and accuracy.

Table 4. Ablation Study on the Hyperparameter α in Our Method PAI: When α becomes excessively large, resulting in an unbalanced response, we terminate the experiment and denote this with a dash (-). F1 values that are considered outliers are highlighted in red.

α	γ	L	**LLaVA**			**Minigpt4**			**Shikra**		
			CHAIR_S	CHAIR_I	F1	CHAIR_S	CHAIR_I	F1	CHAIR_S	CHAIR_I	F1
-	-	-	46.2	13.8	75.9	31.6	10.5	69.4	56.2	15.8	74.6
0.1	1.1	✓	47.4	13.8	76.9	25.0	8.0	69.9	56.4	15.0	75.7
0.2	1.1	✓	47.4	13.6	76.9	**21.2**	**7.7**	70.3	57.2	15.6	75.2
0.4	1.1	✓	42.4	12.4	76.9	15.1	6.3	67.4	55.2	14.4	75.3
0.5	1.1	✓	**24.6**	**6.8**	74.7	5.4	2.9	61.5	51.6	13.2	75.8
0.6	1.1	✓	7.8	4.8	62.6	-	-	-	**37.2**	**10.6**	76.7
0.7	1.1	✓	-	-	-	-	-	-	14.4	6.1	64.9



Please describe the image in detail.

The image depict a kitchen, along with a refrigerator located on the left side of the room. A dishwasher can be seen in the background, and a toaster is placed on the countertop. There are several bottles scattered throughout the kitchen, with some placed on the countertops and others on the floor. A bowl is also visible on the countertop, adding to the overall tidy appearance of the kitchen.

Please describe the image in detail.

The image shows a small, clean kitchen with a window above the sink. The kitchen is well-maintained and organized, featuring a sink, a refrigerator, and a stove. The sink is located under the window, and the refrigerator is situated on the right. The kitchen is well-maintained and appears to be in a hotel or motel setting. The room is clean and well-organized, making it an inviting space for guests.

Please describe the image in detail.

The image features a herd of zebras walking through a grassy field. There are nine zebras in total, spread across the field, with some closer to the foreground and others further in the background…

Please describe the image in detail.

The image features a herd of zebras in a grassy field. There are six zebras in the scene, with some of them standing close to each other, while others are standing further apart…

Figure 7. Illustration of hallucination correction by our proposed PAI with two samples. The left panel shows the response from LLAVA, while the right panel presents the response from LLAVA when integrated with PAI. The descriptions containing hallucinations are emphasized in red.

This makes the comparison with the CHAIR metric fairer when the F1 scores are similar. In the above, the F1 scores of the various methods differ slightly, so we have not included this somewhat redundant measure. The F1 score is calculated by counting the objects included in the description, the objects in the ground-truth sets, and the hallucinated objects.

Table 5. Ablation Study on the Hyperparameter γ and Layer Prior “L” in Our Method PAI: Results are presented for the models LLAVA (left), Minigpt4 (middle), and Shikra (right).

α	γ	L	CHAIR_S	CHAIR_I	F1	α	γ	L	CHAIR_S	CHAIR_I	F1	α	γ	L	CHAIR_S	CHAIR_I	F1
-	-	-	46.2	13.8	75.9	-	-	-	31.6	10.5	69.4	-	-	-	56.2	15.8	74.6
0.5	1.1	✓	24.6	6.8	75.7	0.2	1.1	✓	21.2	7.7	70.3	0.6	1.1	✓	**37.2**	10.6	76.7
0.5	1.2	✓	**23.4**	**6.6**	75.7	0.2	1.2	✓	16.4	**6.2**	70.3	0.6	1.2	✓	40.2	10.5	76.0
0.5	1.3	✓	24.6	7.5	74.4	0.2	1.3	✓	**14.4**	6.9	69.5	0.6	1.3	✓	40.6	10.5	76.0
0.5	1.5	✓	25.0	9.6	74.6	0.2	1.5	✓	12.6	5.4	67.1	0.6	1.5	✓	40.2	10.5	75.3
0.5	2.0	✓	24.2	7.1	74.0	0.2	2.0	✓	8.4	4.6	64.3	0.6	2.0	✓	39.4	**10.1**	76.2
0.5	1.1	✓	**24.6**	**6.8**	75.7	0.2	1.1	✓	**21.2**	**7.7**	70.3	0.6	1.1	✓	**37.2**	**10.6**	76.7
0.5	1.1	×	20.6	6.7	71.7	0.2	1.1	×	22.3	8.4	70.3	0.6	1.1	×	56.4	15.7	74.9

5.4.1 Effects of α in Exciting Image Attention.

In the process of exciting the attention values of image tokens, we introduce a parameter α to control the amplification scale. As shown in Tab. 4, different LVLMs exhibit varying sensitivity to the amplification scale. This sensitivity not only depends on the length of the model’s image tokens (e.g., the image token length of LLaVA-1.5 is 576, while that of Minigpt4 is only 32), but also on the model’s pre-training method.

However, a commonality across these models is that an appropriate amplification scale can achieve a balance between the number of hallucinated objects in the description and the amount of information conveyed. If the scale is too small, the description may still contain many hallucinated objects. Conversely, if the scale is too large, the amount of information in the response will decrease.

5.4.2 Effects of Layer Prior in Exciting Image Attention.

We further investigate the control of the intervention layer for exciting image attention. As observed in Tab. 5, the introduction of this prior does indeed improve the performance of our method. However, when there is no control over the intervention layer and interventions are applied to all layers, the performances of different models exhibit some variations. For Shikra, in the absence of the layer prior, our method’s results regress to the baseline. For both LLaVA and Minigpt4, the loss of the layer prior causes some unfavorable fluctuations in both the CHAIR metric and the F1 score.

5.4.3 Effects of γ in Mitigating Language Prior.

Tab. 5 presents the results of an ablation study focusing on γ, which adjusts the balance between output distributions from conditioned inputs with excited image tokens and pure text inputs. Unlike the other two models, the Minigpt4 model is highly sensitive to γ. When γ is too large, it can lead to uncontrollable model behavior. The experimental results suggest that maintaining γ within a relatively small range, such as 1.1 - 1.2, yields the most stable performance.

6. Conclusion and Limitation

In this paper, we first analyze the causes and manifestations of hallucinations in LVLMs. We propose a phenomenon termed ”text inertia,” wherein the model continues to produce the same hallucinatory descriptions even when no image input is provided. This issue fundamentally stems from the model’s neglect of image tokens. Consequently, we introduce the PAI method to intervene in the model’s inference process, steering it towards an image-based and trustworthy direction. This is a training-free method and does not require any external tools. Extensive experiments on multiple benchmarks and LVLMs have validated the effectiveness of PAI in mitigating hallucination issues.

Limitations: (1) The language decoders of existing open-source LVLMs are primarily models from the LLaMA-family. It is worth exploring whether the issues of image neglect and text inertia are introduced by LLaMA. (2) As we described, our method fundamentally alleviates the image neglect issue during model inference. Its upper limit depends on the capabilities of the well-trained model. Therefore, it is worth investigating whether incorporating this issue as a loss during the training process could lead to further performance improvements.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020. 14

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan.

Flamingo: a visual language model for few-shot learning. 3

[3] Gunjal Anisha, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. Aug 2023. 3

[4] Amos Azaria and Tom Mitchell. The internal state of an llm knows when its lying. Apr 2023. 4

[5] T.B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Askell Amanda, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Henighan Tom, Rewon Child, A. Ramesh, DanielM. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, EricJ. Sigler, Mateusz Litwin, Scott Gray, Chess Benjamin, Jack Clark, Christopher Berner, McCandlish Sam, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *arXiv: Computation and Language,arXiv: Computation and Language*, May 2020. 3

[6] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. Dec 2022. 4

[7] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 782–791, 2021. 14

[8] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, Rui Zhao, Sensetime Research, Sklsde Sklsde, Beihang Beihang, and Seiee Seiee. Shikra: Unleashing multimodal llm's referential dialogue magic. 3

[9] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. Dec 2023. 2, 3

[10] Wenliang Dai, Junnan Li, Dongxu Li, AnthonyMeng Huat, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. 3

[11] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023. 2, 5

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North*, Jan 2019. 4

[13] Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. Chartllama: A multimodal llm for chart understanding and generation. Nov 2023. 1

[14] Hongyu Hu, Jiyuan Zhang, Minyi Zhao, and Zhenbang Sun. Ciem: Contrastive instruction evaluation method for better instruction tuning. 3

[15] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multimodal large language models via over-trust penalty and retrospection-allocation. Nov 2023. 3, 6

[16] Jushi Kai, Tianhang Zhang, Hai Hu, and Zhouhan Lin. Sh2: Self-highlighted hesitation helps you decode more truthfully. Jan 2024. 4

[17] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, Jiaya Jia, Hong Kong, and Microsoft Research. Lisa: Reasoning segmentation via large language model. 1, 3

[18] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. 3, 6

[19] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. 3, 4

[20] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36, 2024. 4, 5

[21] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 2

[22] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. Nov 2023. 2, 3

[23] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2023. 3

[24] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. Sep 2023. 3

[25] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 3

[26] Haotian Liu, Chunyuan Li, Qingyang Wu, YongJae Lee, Madison Madison, and Microsoft Research. Visual instruction tuning. 3

[27] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. 2

[28] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Jan 2022. 3

[29] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. 3

[30] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. 3

[31] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Jan 2018. 2

[32] Guillaume Sanchez, Honglu Fan, Alexander Spangher, Elad Levi, Pawan Sasanka Ammanamanchi, and Stella Biderman. Stay on topic with classifier-free guidance. *arXiv preprint arXiv:2306.17806*, 2023. 5

[33] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023. 2

[34] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented rlhf. Sep 2023. 2, 3

[35] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timoth'ee Lacroix, Baptiste Rozi'ere, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. 3

[36] Alex Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023. 4

[37] Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. Label words are anchors: An information flow perspective for understanding in-context learning. *arXiv preprint arXiv:2305.14160*, 2023. 5

[38] Fei Wei, Xinyu Zhang, Ailing Zhang, Bo Zhang, and Xiangxiang Chu. Lenna: Language enhanced reasoning detection assistant. Dec 2023. 1

[39] Jason Wei, Maarten Bosma, VincentY. Zhao, Kelvin Guu, AdamsWei Yu, Brian Lester, Nan Du, AndrewM. Dai, and QuocV. Le. Finetuned language models are zero-shot learners. *Learning,Learning*, Sep 2021. 3

[40] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. Sep 2023. 5

[41] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-owl: Modularization empowers large language models with multimodality. 3

[42] Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. 3

[43] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. Oct 2023. 3

[44] Yuechen Zhang, Shengju Qian, Bohao Peng, Shu Liu, and Jiaya Jia. Prompt highlighter: Interactive control for multi-modal llms. *arXiv preprint arXiv:2312.04302*, 2023. 1, 13

[45] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. Oct 2023. 3

[46] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. 3

Appendix

A. Text Inertia Detection Process

To detect the text inertia phenomenon in LVLMs during image description tasks, we approach this issue in two stages. In the first stage, we employ the CHAIR metric to identify the indices of hallucinated objects in the LVLMs' descriptions. In the following stage, we extract the tokens preceding the hallucinated objects, feed them into the LVLMs, and proceed with generation, excluding the image from the input. In other words, we only provide the LVLMs with the task instruction and history response up to the index of the hallucinated object. Finally, we extract the first ten tokens from the newly generated output and use GPT-4 to determine whether the same object has been generated. Our prompt structure is shown in Tab. S1.

Table S1. The Prompt Used for Text Inertia Detection.

GPT-4 Prompt

Please assist me in determining whether the following descriptions include the specified object. Simply respond with "Yes" or "No". Consider synonyms and similar expressions.
 	
[Object] {object}
[Description] {description}

Here, {object} is replaced by the hallucinated object that we extract each time, and {description} is replaced by the first ten tokens of the newly generated conditioned description. When the model answers "Yes", it indicates the presence of the text inertia phenomenon.

Table S2. Performance of Our Method PAI on the LLaVA-1.5 (13B) Model. We present the average Accuracy and F1 score across the three splits of POPE.

Model	**Method**	**CHAIR**			**POPE**	
		CHAIR_S	CHAIR_I	F1	Acc	F1
LLaVA-1.5(13B)	Greedy	44.0	12.7	77.3	85.47	86.60
	PAI	**33.0**	**9.2**	77.8	**86.82**	**87.80**

B. What if Model Scales Up?

We performed experiments on the LLaVA-1.5-13B scale model to assess how the performance of PAI varies with increasing model scale. As shown in Tab. S2, our method maintains its robustness regardless of the scale of the model. Notably, it continues to alleviate the hallucination issue in both long sequence and short VQA tasks as the model scale increases.

Table S3. Comparison with Different Intervention Method. We present the average Accuracy and F1 score across the three splits of POPE.

Model	**Method**	**CHAIR**			**POPE**	
		CHAIR_S	CHAIR_I	F1	Acc	F1
LLaVA-1.5(7B)	Greedy	46.2	13.8	75.9	84.76	85.59
	PH	52.8	14.2	76.9	84.11	85.20
	PAI	**23.4**	**6.6**	75.7	**86.13**	**86.42**
LLaVA-1.5(13B)	Greedy	44.0	12.7	77.3	85.47	86.60
	PH	46.6	12.8	77.3	85.69	86.78
	PAI	**33.0**	**9.2**	77.8	**86.82**	**87.80**

C. Comparative Experimental Results

We compare our method with a different intervention approach during model inference. **Prompt Highlighter** [44] proposes an interactive technique for constructing visual prompts to steer model generation. Essentially, it adds a value of the same size to the attention mask of the tokens in the user's guidance area to influence the model's output. To automate the intervention process for evaluation, we adopt an approach similar to Prompt Highlighter. This involves adding a constant value of 1.0 to the attention mask of all image tokens, where 1.0 is the default value provided by their open-sourced code.

The final experimental results are presented in Tab. S3. The findings indicate that uniformly applying intervention across the entire image is less effective than enhancing image attention based on the original scale.

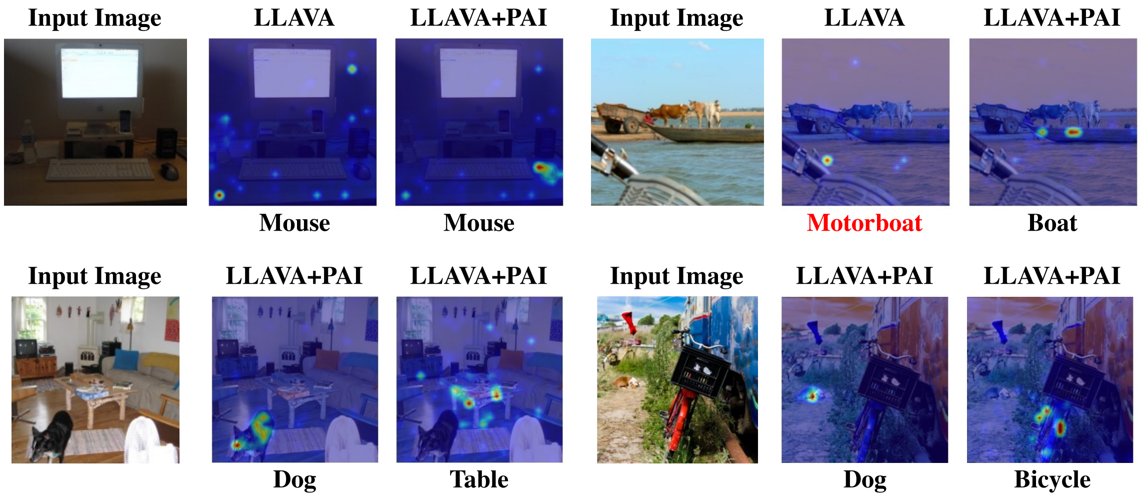


Figure S1. Visualization of the self-attention maps for each object token with and without PAI over LLAVA. The hallucinated objects are higlighted in red.

D. Obtaining Explainable LVLMs.

The extent to which attention can serve as an explanatory mechanism has been extensively studied [1, 7]. Within the scope of LVLMs, the self-attention maps in the LLaMA decoder have been regarded as a natural explanation for the model. However, these explanatory results are only available for LLAVA and Shikra, due to the usage of linear projection that preserves all image patch token features. The model, which employs a resampler to derive the explanation, has to reverse-engineer the abstracted image token back to the input image patch token. However, tracing back the image patch after it has been encoded by the visual encoder and resampled into an image token proves to be challenging.

As illustrated in Fig. S1, the original LLAVA identifies the correct object, a mouse, and hallucinates an object, a motorboat, both of which have limited interpretability as the results do not focus on the main body of the described object. With our method of layer intervention, the response becomes more image-centric. The description aligns better with the image location, which can all be considered as a more faithful explanation.

Table S4. The prompt used for GPT-4V evaluation

GPT-4V Prompt

You are required to score the performance of two AI assistants in describing a given image. You should pay extra attention to the hallucination, which refers to the part of descriptions that are inconsistent with the image content, such as claiming the existence of something not present in the image or describing incorrectly in terms of the counts, positions, or colors of objects in the image. Please rate the responses of the assistants on a scale of 1 to 10, where a higher score indicates better performance, according to the following criteria:
1: Accuracy: whether the response is accurate with respect to the image content. Responses with fewer hallucinations should be given higher scores.
2: Detailedness: whether the response is rich in necessary details. Note that hallucinated descriptions should not count as necessary details.
Please output the scores for each criterion, containing only two values indicating the scores for Assistant 1 and 2, respectively. The two scores are separated by a space. Following the scores, please provide an explanation of your evaluation, avoiding any potential bias and ensuring that the order in which the responses were presented does not affect your judgment.

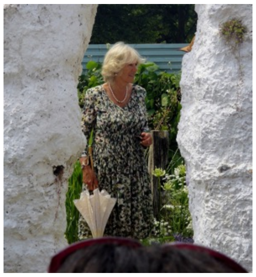
[Assistant 1]
{Response of Assistant 1}
[End of Assistant 1]

[Assistant 2]
{Response of Assistant 2}
[End of Assistant 2]

Output format:
Accuracy: <Scores of the two answers>
Reason:

Detailedness: <Scores of the two answers>
Reason:

Instruction: Please describe the image in detail.



Input Image

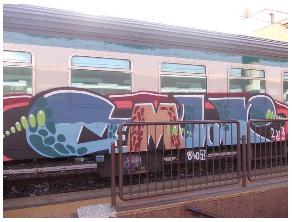
LLaVA-1.5(Greedy):
The image features a woman wearing a floral dress, walking through a garden with a parasol in her hand. She appears to be enjoying her time in the garden, possibly posing for a picture. A handbag can also be seen in the scene, placed close to the woman, possibly belonging to her. The overall atmosphere of the image is pleasant and inviting, showcasing the beauty of the garden.

LLaVA-1.5(PAI):
The image features a smiling woman wearing a floral dress, walking through a garden. She is holding an umbrella in her hand, and appears to be enjoying her time in the garden. The woman is walking between two large white pillars, which are part of a garden wall. The scene is set in a garden, with a potted plant nearby. The woman's smile and the pleasant atmosphere of the scene suggest a pleasant and relaxing environment.

Instruction: Please describe the image in detail.

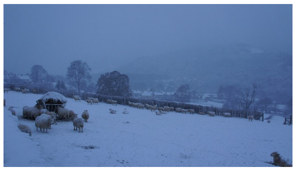
LLaVA-1.5(Greedy):
The image features a train with graffiti on its side, traveling down the tracks. The train is positioned on the left side of the image, and the graffiti is visible on the side of the train, covering a significant portion of the surface. There is a person standing near the train, possibly observing or waiting for it to stop. The scene also includes a bench located on the right side of the image, providing a place for people to sit and wait for the train.

LLaVA-1.5(PAI):
The image features a train with graffiti on its side, which is parked at a train station. The train is blue and white in color, and it is parked on the tracks. The graffiti on the train is quite elaborate, covering a significant portion of the train's side. The train is parked on the tracks, and it appears to be a passenger train, as it is not a freight train. The train is parked at a station, and it is likely waiting for passengers to board. The graffiti on the train adds a unique and artistic touch to the otherwise ordinary scene.



Input Image

Instruction: Please describe the image in detail.



Input Image

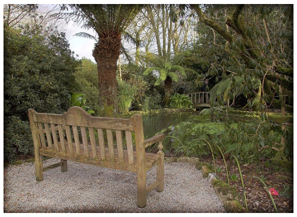
LLaVA-1.5(Greedy):
The image depicts a snowy hillside with a herd of sheep grazing on the grass. There are at least 14 sheep scattered across the field, with some closer to the foreground and others further away. The sheep are of various sizes, indicating a mix of adult and younger animals. The scene captures the beauty of the snow-covered landscape and the peacefulness of the animals as they graze in the wintery environment.

LLaVA-1.5(PAI):
The image features a large herd of sheep standing in a snowy field. There are at least 14 sheep in the scene, with some of them standing close to each other, while others are spread out across the field. The sheep are standing in the snow, and the scene appears to be a peaceful, snowy landscape. The image captures the beauty of the winter landscape and the calmness of the sheep in their natural environment.

Instruction: Please describe the image in detail.

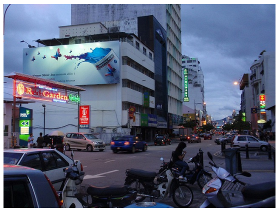
LLaVA-1.5(Greedy):
The image features a wooden bench situated in a serene garden setting. The bench is positioned near a tree, providing a shaded and relaxing spot for visitors. The garden is adorned with several potted plants, creating a lush and green atmosphere. In addition to the bench, there are two people in the scene, one standing near the left side of the bench and the other closer to the right side. The presence of these individuals adds a sense of life and activity to the otherwise peaceful garden setting.

LLaVA-1.5(PAI):
The image features a wooden park bench situated in a park-like setting, surrounded by a variety of trees and plants. The bench is located in the middle of the scene, with a few trees and bushes in the background. The bench is surrounded by a few bushes and trees, creating a serene and peaceful atmosphere. The bench is situated in the middle of the scene, with a few trees and bushes in the background. The bench is made of wood and is placed on a gravel surface, adding to the overall charm of the scene.



Input Image

Figure S2. PAI's performance on reducing hallucinations of LLaVA-1.5-7B.



Input Image

Instruction: Please describe the image in detail.

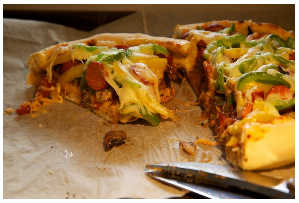
Minigpt4(Greedy):
This image shows a busy street with cars parked on both sides of the road. The buildings on either side of the street are tall and white, with large billboards on the sides. The sky is cloudy and dark, with a few dark clouds visible in the distance. The street is lined with trees and streetlights.

Minigpt4(PAI):
This image shows a busy city street with cars parked on the side of the road. The building in the background is a tall white building with a large billboard on the side. The billboard has a blue and white design on it. There are also several motorcycles parked on the side of the road.

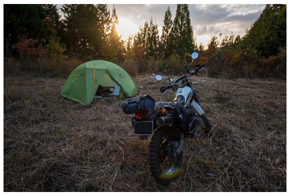
Instruction: Please describe the image in detail.

Minigpt4(Greedy):
This image shows a pizza with various toppings, including peppers, onions, and mushrooms, on a piece of parchment paper. The pizza appears to be cut into slices, with some slices missing. There is a knife and fork next to the pizza, and a bottle of tomato sauce on the table in the background.

Minigpt4(PAI):
This image shows a pizza that has been cut into slices and is sitting on a piece of parchment paper. The pizza has a variety of toppings, including peppers, onions, and mushrooms, and is covered in a thick layer of cheese. A knife is visible in the foreground, next to the pizza.



Input Image



Input Image

Instruction: Please describe the image in detail.

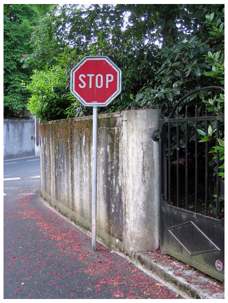
Minigpt4(Greedy):
The image shows a motorcycle parked in a field next to a tent. The tent is set up on a wooden platform and has a green tarp covering it. The motorcycle is parked on the grass and has a helmet on the handlebars. There are trees in the background and the sun is setting behind them. The sky is orange and pink, and there are clouds in the sky.

Minigpt4(PAI):
This image shows a motorcycle parked in a field next to a tent. The tent is set up on a wooden platform and has a green tarp covering it. There are trees in the background and the sun is setting behind them. The motorcycle is parked on the ground next to the tent, and there are no other objects in the image.

Instruction: Please describe the image in detail.

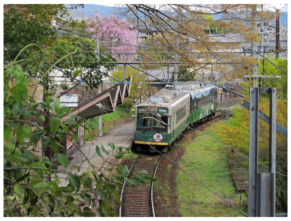
Minigpt4(Greedy):
This image shows a stop sign on the side of a road, next to a brick wall. The stop sign has a red background with white letters that read "stop." There is a gate in the wall, with a black iron fence and a wooden door. The gate is closed, and there are no people or cars in the image. The trees on the other side of the road are in full bloom, with pink flowers covering the branches. The sky is clear and blue, with a few clouds in the distance.

Minigpt4(PAI):
This image shows a stop sign on the side of a road, with a gate in the background. The stop sign has a red background with white letters that read "stop." The gate in the background is made of metal and has a black finish. There are some leaves on the ground next to the stop sign, and there are trees in the background.



Input Image

Figure S3. PAI's performance on reducing hallucinations of Minigpt4.



Input Image

Instruction: Please describe the image in detail.

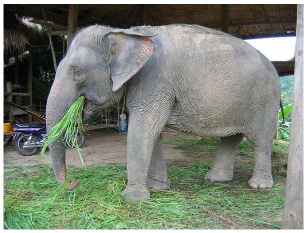
Shikra(Greedy):
The image features a green and yellow train traveling down the tracks, passing by a station. The train appears to be a trolley-type train that is either leaving the station or pulling into it.
There are several people around the train and the station, possibly waiting for the train or just going about their day.
In addition to the train and people, there are a few traffic lights scattered around the scene. One traffic light is located near the top right corner, another one is positioned further to the right, and the third traffic light is on the left side of the image. The presence of these traffic lights suggests that the train is operating in an urban environment with well-organized transportation infrastructure.

Shikra(PAI):
The image depicts a green and yellow train traveling down the train tracks, passing through a small town. The train is moving along the tracks that are situated above a grassy field. The scene also features a tree-lined sidewalk adjacent to the train tracks.
In the background, there are some pink flowers in bloom, adding a touch of color to the scene. The train tracks can be seen extending across the image, with the train positioned in the middle of the frame. The train appears to be a local commuter train, possibly heading towards a city or town.

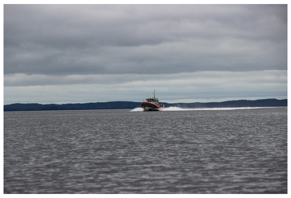
Instruction: Please describe the image in detail.

Shikra(Greedy):
The image features a large elephant standing on a lush green field, eating grass using its trunk. The elephant is positioned in the center of the scene, and it appears to be enjoying a meal.
In the background, there are several chairs scattered around the field, possibly for people to sit and observe the elephant. Additionally, there are two people in the scene, one near the left edge of the image and another closer to the center, both of them likely watching the elephant as it feeds.

Shikra(PAI):
The image features a large elephant standing on a dirt field, eating a mouthful of green grass. The elephant is enjoying a meal consisting mainly of grass, as it is eating the grass in its mouth. The elephant is standing under a bamboo hut, which provides shade and shelter.\n\nThe image shows a large elephant standing in a field, eating grass by placing it in its mouth using its trunk. The scene captures the unique moment of the elephant enjoying its meal.



Input Image



Input Image

Instruction: Please describe the image in detail.

Shikra(Greedy):
The image features a small boat sailing across a large body of water, likely an ocean or a sea, under a cloudy sky. The boat is speeding through the water, cutting through the waves and creating a wake behind it.
There are a few people on the boat, with one person visible on the left side and two other people on the right side of the boat. They seem to be enjoying their time on the water, possibly engaged in recreational activities or sightseeing.

Shikra(PAI):
The image features a small red boat, resembling a red and white ship, speeding across the water with a large body of waves around it. The boat is moving quickly through the water, creating a wake in the ocean. The scene captures the excitement of an adventurous boat ride on the open ocean, providing a thrilling experience for those on board.

Figure S4. PAI's performance on reducing hallucinations of Shikra.