

---

# On Epistemic Uncertainty of Visual Tokens for Object Hallucinations in Large Vision-Language Models

---

Hoigi Seo<sup>1\*</sup> Dong Un Kang<sup>1\*</sup>  
Hyunjin Cho<sup>1</sup> Joohoon Lee<sup>2</sup> Se Young Chun<sup>1,2,3†</sup>  
<sup>1</sup>Dept. of ECE <sup>2</sup>IPAI & <sup>3</sup>INMC, Seoul National University, Republic of Korea  
{seohoiki3215, qkrtnskfk23, jim0228, joohoon1, sychun}@snu.ac.kr

## Abstract

Large vision-language models (LVLMs), which integrate a vision encoder (VE) with a large language model, have achieved remarkable success across various tasks. However, there are still crucial challenges in LVLMs such as object hallucination, generating descriptions of objects that are not in the input image. Here, we argue that *uncertain* visual tokens within the VE is a key factor that contributes to object hallucination. Our statistical analysis found that there are positive correlations between visual tokens with high *epistemic* uncertainty and the occurrence of hallucinations. Furthermore, we show theoretically and empirically that visual tokens in early VE layers that exhibit large representation deviations under small adversarial perturbations indicate high epistemic uncertainty. Based on these findings, we propose a simple yet effective strategy to mitigate object hallucination by modifying the VE only. Our method comprises a proxy method with adversarial perturbations for identifying uncertain visual tokens efficiently and a method to mask these uncertain visual tokens during the self-attention process in the middle layers of the VE, suppressing their influence on visual encoding and thus alleviating hallucinations. Extensive experiments show that our method significantly reduces object hallucinations in LVLMs and can synergistically work with other prior arts.

## 1 Introduction

Large Vision-Language Models (LVLMs) have demonstrated impressive capabilities across a range of multi-modal tasks, including image captioning [1, 10, 30, 33, 63], visual-question answering (VQA) [10, 43, 58], and multi-modal dialogue systems [14, 28, 35, 36, 69]. Despite these notable advancements, recent studies [19, 31, 47, 57] have reported that LVLMs are susceptible to hallucination, generating textual descriptions that do not align with the input image. In particular, object hallucination, where the model describes objects not present in the input image, significantly undermines the reliability and thus the practical utility of LVLMs [20, 24, 25, 27].

To mitigate object hallucination in LVLMs, recent works [2, 8, 20, 24, 25, 27, 37] have explored training-free approaches including modifying the decoding strategy of the language model [2, 20, 27, 37], modulating attention mechanisms [24, 25, 37], or altering the input image [2] during inference. While these methods have shown effectiveness in reducing object hallucination, they often suffer limitations such as requiring multiple inferences of the large language model, which is the most computationally expensive component of LVLMs, or yielding relatively small performance gains. In contrast, approaches for object hallucination mitigation that directly target the vision encoder, a core component responsible for visual perception, have been relatively underexplored.

---

\* equal contribution, † corresponding author.

In this work, we investigate how visual information contributes to object hallucination in LVLMs, with a particular focus on the uncertainty of visual tokens introduced by the pre-trained vision encoder (*i.e.*, epistemic uncertainty). Estimating this uncertainty typically requires intensive computation, such as Monte Carlo (MC) dropout [44], which involves thousands of forward passes. To provide a more efficient alternative, we present a theoretical analysis showing that the deviation of visual token representations under adversarial perturbations is monotonically related to an upper bound of uncertainty for each visual token, particularly in the early layers of the vision encoder. Empirically, we find that the norm of representation deviation in visual tokens caused by adversarial perturbations closely aligns with uncertainty estimates obtained via MC dropout, enabling a more efficient approximation of visual token uncertainty. Furthermore, we empirically demonstrate a strong positive correlation between visual token uncertainty and the occurrence of object hallucination of LVLMs.

Motivated by this observation, we propose a simple yet effective method to mitigate hallucination by intervening only in the vision encoder during inference. Specifically, we first identify *uncertain visual tokens*, defined as those whose representations exhibit significant deviation under PGD-based adversarial perturbations [40] which reflect high epistemic uncertainty. We then suppress their influence by masking these uncertain tokens in the self-attention layers of intermediate vision encoder blocks. This approach reduces the model’s dependence on uncertain visual features while preserving the global semantic structure of the image representation.

Extensive experiments demonstrate that our method effectively reduces object hallucination on benchmark datasets such as CHAIR [47], POPE [31], and AMBER [56]. We validate our approach across a range of LVLM architectures [9, 35, 69], incorporating diverse vision encoders, language models, and training regimes to ensure generalizability. Notably, because our method exclusively modifies the vision encoder, it can be seamlessly combined with existing methods that adjust decoding strategies or attention mechanisms within the language model.

Our contribution can be summarized as follows.

- We theoretically and empirically demonstrate that the visual tokens exhibiting the representation deviations under adversarial perturbations indicate upper bound of epistemic uncertainty, which is strongly correlated with object hallucination in LVLMs.
- Motivated by this insight, we propose an efficient and effective method that mitigates hallucination by identifying uncertain visual tokens via adversarial perturbation and masking them in the self-attention layers of intermediate vision encoder blocks.
- Our method is validated across multiple benchmarks and LVLM architectures, and is easily compatible with existing mitigation methods, enabling synergistic gains in performance.

## 2 Related Works

**Large Vision-Language Models.** Large Vision-Language Models (LVLMs) integrate visual and textual inputs for multi-modal reasoning and generation. Modern LVLMs typically consist of a vision encoder [15, 17, 22, 46, 65], a connector, and a language model [3, 11, 54, 62]. Some use linear projections to align visual features with the language embedding space [9, 36], while others adopt Q-Former modules [14, 30, 69] that use learnable queries to extract and compress visual information. Despite their remarkable performance on multi-modal tasks, LVLMs exhibit hallucination, generating output misaligned with visual content, raising concerns about their reliability in real-world usage.

**Mitigating hallucinations in LVLMs.** Hallucination in LVLMs refers to the phenomenon in which the output contradicts the visual input by fabricating visual information [5, 34]. Mitigation strategies fall into training-based and training-free categories. Training-based methods optimize the LVLMs [23, 64] or incorporate auxiliary modules for output guidance [16, 39, 68], but are often computationally expensive. Training-free approaches modify logits of language models to suppress hallucination-prone text tokens [2, 20, 21, 27, 29, 37, 55, 70], adjust attention process [24, 25, 29, 37, 60], or modify inputs [2, 42, 66]. However, the approaches overlook deficiencies in the vision encoder. We instead propose an orthogonal and training-free strategy: leverage adversarial attacks to identify uncertain visual tokens and suppress them, complementing language-level approaches.

**Adversarial attack on LVLMs.** Adversarial attack [18, 40, 53] introduces imperceptible perturbations in images to induce incorrect predictions by a model. While early efforts focused on tasks

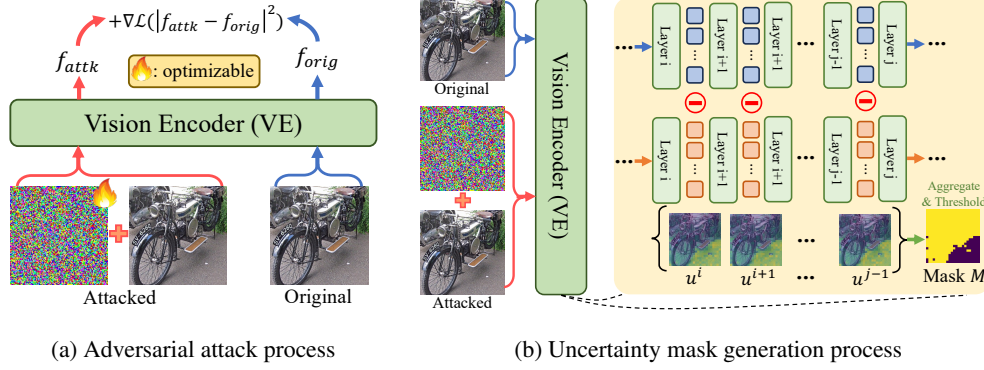


Figure 1: **Overall illustration of the adversarial attack and uncertainty mask generation process.** (a) The original image is processed by the vision encoder (VE) to obtain features  $f_{\text{orig}}$ . An adversarial image is created by adding optimizable noise, which is then encoded to produce  $f_{\text{attk}}$ . The noise is optimized using Projected Gradient Descent (PGD) to maximize the mean squared error between  $f_{\text{orig}}$  and  $f_{\text{attk}}$ , as described in Eq. 1. (b) From layers  $i$  to  $j-1$ , we extract feature sets  $\mathcal{F}_{\text{orig}} = \{f_{\text{orig}}^i, \dots, f_{\text{orig}}^{j-1}\}$  and  $\mathcal{F}_{\text{attk}} = \{f_{\text{attk}}^i, \dots, f_{\text{attk}}^{j-1}\}$ . The norm differences of corresponding features form layer-wise uncertainty maps  $\mathcal{U} = \{u^i, \dots, u^{j-1}\}$ . These maps are min-max normalized, aggregated, and standardized to produce the final binary uncertainty mask  $M$  using a threshold  $\sigma_{\text{th}}$ .

such as classification [18, 41] and object detection [6, 32, 61], recent work has extended attacks to LVLMs [7, 45, 49, 50, 67] to improve the robustness of the models. In image-targeted attacks, where input is in a discrete pixel space, Projected Gradient Descent (PGD) [40] remains a dominant strategy due to its effectiveness. The optimization process of PGD is formalized as follows.

$$\hat{x}_{i+1} = \Pi\left(\hat{x}_i + \alpha \cdot \text{sign}(\nabla_{\hat{x}_i} \mathcal{L}(F(\hat{x}_i), F(x)))\right), \quad (1)$$

where  $\alpha \in \mathbb{N}$  is the learning rate,  $F$  is a target neural network,  $x$  is the original image,  $\hat{x}_i$  denotes the perturbed image at iteration  $i$ , and  $\Pi$  projects onto the constraint set  $\|\hat{x}_{i+1} - x\|_{\infty} \leq k$ . LVLMs show strong multi-modal capabilities but remain vulnerable to adversarial attacks [7, 45, 59, 67], which can target the entire model [7, 45, 67] or specifically the vision encoder [59].

### 3 Method

In this section, we present our approach for identifying uncertain visual tokens within the vision encoder using adversarial perturbations, as detailed in Sec.3.1. We demonstrate that these tokens significantly contribute to object hallucination in LVLMs through statistical analysis. Based on these findings, we propose a masking strategy within the vision encoder to suppress the influence of uncertain tokens, resulting in a notable reduction in hallucinations, as described in Sec.3.2.

#### 3.1 Adversarial Attack to Vision Encoder Reveals Uncertain Visual Tokens

##### 3.1.1 Efficient uncertainty approximation of visual token with adversarial attack

Estimating uncertainty induced by deep neural networks (*i.e.* epistemic uncertainty) is commonly approached by approximating Bayesian inference using Monte Carlo (MC) dropout [26, 44]. However, the approximation process introduces substantial overhead as a result of thousands of forward passes. In this work, we find that the epistemic uncertainty of individual visual tokens differs from each other, as perceived by the vision encoder for a given image, and their upper bound can be efficiently estimated via adversarial attacks. To support this claim, we first introduce the following lemma.

**Lemma 3.1** (Approximate local Gaussianity under small perturbation). *Let  $f = \{f_t\}_{t=1}^L$  be a smooth  $L$ -layer neural network parameterized by  $\theta$ . For an input  $x \in \mathbb{R}^{N \times 3}$ , define the hidden state at layer  $t$  as  $z^{(t)} = f_t \circ \dots \circ f_1(x)$ . For a perturbed input  $x + \epsilon$ , with  $\|\epsilon\|_{\infty} \leq k$  for sufficiently small  $k > 0$ , define the perturbed hidden state as  $Z^{(t)} = f_t \circ \dots \circ f_1(x + \epsilon)$ . Then, under the assumption that the perturbation is small and  $f \in C^2$ ,  $Z^{(t)}$  can be locally approximated by a Gaussian centered at  $z^{(t)}$ , with a third-order remainder in the log-density.*

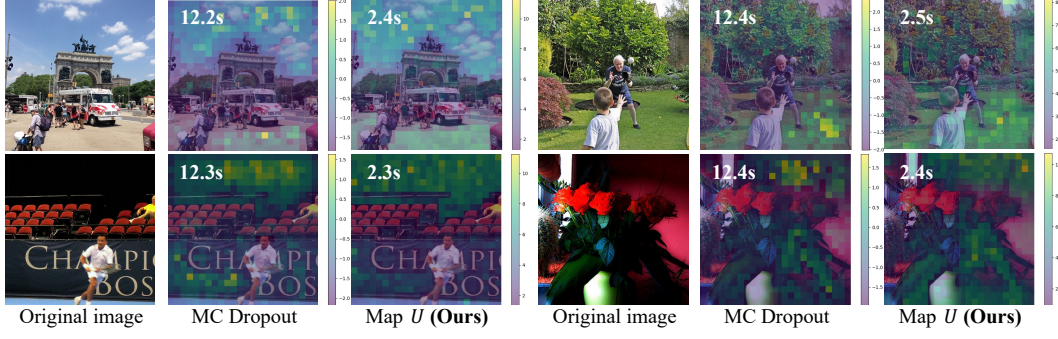


Figure 2: **Visual comparison of estimated uncertainty from MC dropout [44] and our method.** Our uncertainty map  $U$  identifies uncertain regions similar to the uncertainty map obtained via MC dropout. MC dropout was applied to the residuals of the LLaVA-1.5 vision encoder with a dropout rate of  $p = 0.5$  and the variance of each token was estimated over 1,000 forward passes. For the adversarial attack, we applied 100 iterations of PGD with  $k = 3$ . The MC-based uncertainty values were log-scaled for visualization clarity. The runtime for each example is shown in the top-left corner.

The proof of Lemma 3.1 can be found in the Appendix Sec. A.1. The lemma implies that the hidden states exhibit Gaussianity under small perturbation, which allows us to prove the following theorem.

**Theorem 3.2** (Upper bound of differential entropy increases as hidden state deviation increases under adversarial attack). *Let  $x$  be an input image, and let  $\epsilon$  be a small adversarial perturbation. Define the perturbed input as  $X := x + \epsilon$ . Let  $f = \{f_t\}_{t=1}^L$  be a smooth  $L$ -block transformer that processes a sequence of  $N$  input tokens. Let  $z^{(t)} := f_t \circ \dots \circ f_1(x) \in \mathbb{R}^{N \times d}$  and  $Z^{(t)} := f_t \circ \dots \circ f_1(X) \in \mathbb{R}^{N \times d}$  be the hidden states at layer  $t$  for the clean and perturbed inputs, respectively. Denote the  $i$ -th token representation at layer  $t$  as  $z_i^{(t)} \in \mathbb{R}^d$  and  $Z_i^{(t)} \in \mathbb{R}^d$ . If  $Z_i^{(t)}$  changes smoothly with small  $\epsilon$ , then the upper bound of the differential entropy of  $Z_i^{(t)}$  increases as  $\mathbb{E}_\epsilon[\|Z_i^{(t)} - z_i^{(t)}\|_2^2]$  increases.*

The proof of Theorem 3.2, provided in Appendix Sec. A.2, shows that under adversarial attack, the norm of hidden state deviation efficiently approximates the upper bound of visual token’s entropy.

Leveraging this insight from Theorem 3.2, we aim to obtain a mask that identifies uncertain visual tokens with an adversarial attack. Specifically, given an image  $x$  and a vision encoder  $F_V$ , we first obtain the feature  $f_{\text{orig.}} = F_V(x) \in \mathbb{R}^{N \times d}$ , where  $N$  denotes the number of image tokens. We then generate an adversarially perturbed image  $\hat{x}_0$  by adding small noise  $\epsilon$  to  $x$  such that  $\|\epsilon\|_\infty \leq k$ . We then extract feature of perturbed image with  $f_{\text{attk.}} = F_V(\hat{x}_0)$ . We define the adversarial objective as the mean squared error between  $f_{\text{orig.}}$  and  $f_{\text{attk.}}$ , and optimize  $\epsilon$  with PGD for  $I$  iterations as specified in Eq. 1 to obtain the final attacked image  $\hat{x} := \hat{x}_I$ . This attack process is illustrated in Fig. 1a.

Next, we extract the hidden states from each layer of the  $F_V$  within the consecutive layer index set  $\mathcal{S} = \{i, \dots, j-1\}$  for both the original image  $x$  and the perturbed image  $\hat{x}$ . This results in two hidden states sets:  $\mathcal{F}_{\text{orig}} = \{f_{\text{orig}}^i, \dots, f_{\text{orig}}^{j-1}\}$  from  $x$  and  $\mathcal{F}_{\text{attk}} = \{f_{\text{attk}}^i, \dots, f_{\text{attk}}^{j-1}\}$  from  $\hat{x}$ . For each layer  $l \in \mathcal{S}$ , we compute the norm of deviation between the corresponding hidden states defined as  $u^l = \|f_{\text{attk}}^l - f_{\text{orig}}^l\|_2$ , resulting in a set of layer-wise uncertainty maps  $\mathcal{U} = \{u^i, \dots, u^{j-1} | \forall l \in \mathcal{S}\}$ .

We then aggregate the layer-wise uncertainty maps in  $\mathcal{U}$  to produce the aggregated uncertainty map  $U$  by applying min-max normalization to each  $u^l$  and averaging across layers, as defined below:

$$U = \frac{1}{j-i} \sum_{l=i}^{j-1} \frac{u^l - u_{\min}^l}{u_{\max}^l - u_{\min}^l}. \quad (2)$$

Finally, we normalize the uncertainty map  $U$  using its mean  $\mu_U$  and standard deviation  $\sigma_U$ , and binarize it with a threshold parameter  $\sigma_{\text{th}}$  to obtain the binary uncertainty mask  $M$  as follows:

$$M = 1 - \frac{1}{2} \left[ \text{sign} \left( \left( \frac{U - \mu_U}{\sigma_U} \right) - \sigma_{\text{th}} \right) + 1 \right] \in \{0, 1\}^N. \quad (3)$$

Here, a value of 0 in the mask  $M$  indicates an “uncertain” visual token, while 1 denotes a relatively “certain” one. Figure 1b illustrates the mask generation process, and examples of  $M$  are shown in Appendix Sec.G.3. In Sec 3.2, we describe how  $M$  is used to mitigate object hallucination.



### 3.1.2 Empirical study on extracting uncertainty with adversarial attack

**Comparison with uncertainty via MC dropout.** We compare our uncertainty map  $U$  with MC dropout [44] to assess how well  $U$  approximates epistemic uncertainty. As shown in Fig.2, the results indicate that  $U$  closely aligns with the uncertainty estimated via MC dropout, demonstrating that  $U$  serves as an efficient approximation. On average, it is approximately 5 times faster than MC dropout in practice. Additional qualitative and computational cost comparisons are provided in Appendix Sec.E.1.

**The range of layer indices set  $\mathcal{S}$  of vision encoder.** As described in Sec. 3.1.1, we extract hidden states from the consecutive layer index set  $\mathcal{S}$ . Our Lemma 3.1 and Theorem 3.2 rely on the assumption that adversarially induced norm of visual feature deviations are small, requiring that perturbations remain limited. Fig.3 shows these deviations are minor in early layers but increase in later ones. To ensure consistency with both the theoretical assumptions and empirical observations, we construct  $\mathcal{S}$  from early layers of vision encoder. Further analyses on  $\mathcal{S}$ , provided in Sec.4.3, additionally support this theoretical and empirical alignment.

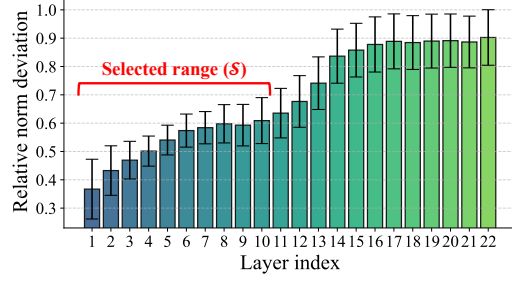


Figure 3: **Relative deviation between attacked and original features.** We used 500 images from the MS-COCO [33] with LLaVA-1.5 vision encoder [35]. Perturbations introduced through the vision encoder remain minimal in early layers but intensify in later ones. We extract the mask from early layers where feature deviations are comparatively small. Error bars denote the  $2\sigma$  range.

## 3.2 Mitigating Object Hallucination of LVLMs via Uncertain Visual Tokens

Building on the identification of uncertain visual tokens through adversarial perturbations in Sec. 3.1, we now investigate how these tokens can be utilized to reduce object hallucination in LVLMs.

### 3.2.1 Uncertain visual tokens contribute to object hallucination

To assess the practical relevance of uncertain visual tokens in object hallucination, we conducted a preliminary study using LLaVA-1.5-7B [35] on 1,000 randomly sampled images from MS-COCO [33]. We estimate the uncertainty map of each visual token via Monte Carlo (MC) dropout, by computing the token-level variance. Using Eq. 3 and a threshold of  $\sigma_{th} = 1$ , we generate an uncertainty mask and calculate the average variance across the uncertain visual tokens in each image. The resulting averaged variances are grouped into 10 bins, and for each bin, we evaluate the severity of hallucination using the CHAIR [47] benchmark.

The experimental results are presented in Fig. 4. Fig. 4 shows that higher average uncertainty of visual tokens corresponds to more severe object hallucination. To statistically validate this monotonic trend, we performed Spearman’s rank correlation analysis between the average uncertainty (measured via token-level variance) and each hallucination metric. The resulting correlation coefficients were  $\rho = 0.794$  ( $p$ -value = 0.006) for CHAIR<sub>s</sub>,  $\rho = 0.733$  ( $p$ -value = 0.016) for CHAIR<sub>i</sub>, and  $\rho = -0.745$  ( $p$ -value = 0.013) for the F1 score, all statistically significant at  $p$ -value < 0.05, and indicating *strong monotonic relationships* [48] ( $|\rho| > 0.7$ ). Through this statistical analysis, we confirm that uncertain visual tokens contribute to hallucination of LVLMs.

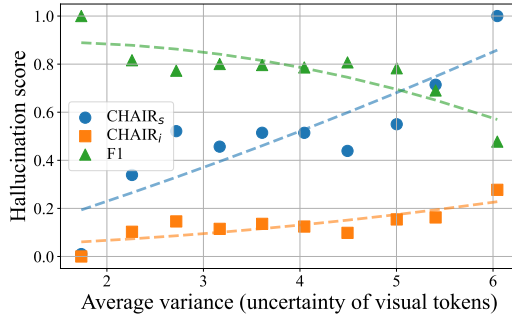


Figure 4: **Relationship between uncertain visual tokens and object hallucination.** The  $x$ -axis represents the average variance within each bin, while the  $y$ -axis shows the corresponding metric scores. The results indicate that higher uncertainty is associated with more object hallucination, with  $p$ -value < 0.05. The trend line was fitted with quadratic function. Note that *higher values* of CHAIR<sub>s</sub> and CHAIR<sub>i</sub>, and *lower* F1 score indicate more *severe object hallucinations*.

### 3.2.2 Masking uncertain visual tokens for training-free hallucination mitigation

Building on the findings in Sec. 3.2.1, we propose a method to reduce object hallucination by leveraging the uncertainty mask  $M$ , which highlights uncertain visual tokens identified through adversarial perturbation. Instead of completely removing these tokens, we attenuate their influence during the self-attention process in the intermediate layers of the vision encoder. The intermediate layers of vision encoder was selected on the basis of empirical evidence that indicates its superior effectiveness in mitigating object hallucination.

Let  $Q, K, V \in \mathbb{R}^{N \times d'}$  be the query, key, and value matrices in a self-attention layer, where  $N$  denotes the number of tokens and  $d'$  the dimensionality of the hidden states within self-attention process. Let  $M \in \{0, 1\}^N$  be the binary uncertainty mask obtained from Eq. 3. Then, our masking strategy modifies the attention computation as follows:

$$\text{Attention}(Q, K, V, M) = \left( \text{Softmax} \left( \frac{QK^\top}{\sqrt{d'}} \right) V \right) \odot M \quad (4)$$

Here,  $\odot$  denotes token-wise multiplication. This operation reduces the influence of uncertain tokens in the attention output while keeping the attention weights and other token interactions intact. Since the masking is applied within the residual connection structure, the model retains stable and meaningful visual representations while suppressing the contribution from uncertain visual tokens. We illustrate this masking strategy within the self-attention process of the vision encoder within LVLMs in Fig. 5.

Compared to masking strategies applied at the input or output of the vision encoder, intervening during self-attention computation in intermediate layers of the vision encoder offers a more balanced approach to reduce the effect of uncertain tokens without discarding potentially useful visual information, as shown in the ablation study in Sec. 4.3.

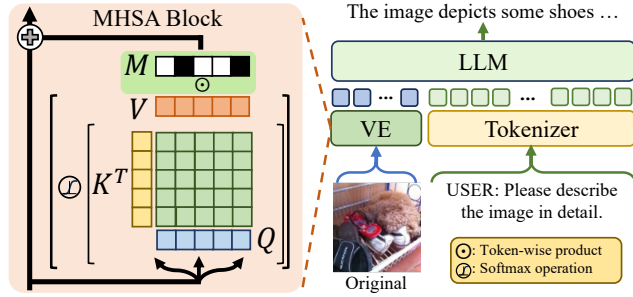


Figure 5: **Illustration of our attention masking method during inference.** In the intermediate multi-head self-attention layers of the vision encoder, we apply a binary uncertainty mask  $M$  to the attention outputs. This token-wise masking reduces the influence of uncertain visual tokens, while preserving the meaningful visual representation.

### 3.2.3 Does our method reduce uncertainty and mitigate object hallucination? Yes.

Based on the relationship between uncertainty of visual tokens and object hallucination discussed in Sec. 3.2.1, we mitigate object hallucination using the method introduced in Sec. 3.2.2. To evaluate the effectiveness of our method in reducing visual token uncertainty, we conducted the same experiment as shown in Fig. 4.

The results in Fig. 6 show that the average variance in the bin with the highest uncertainty decreases from 6.04 to 4.98.  $\text{CHAIR}_s$  drops from 1.00 to 0.33,  $\text{CHAIR}_i$  from 0.27 to 0.09, and the F1 score increases from 0.47 to 0.77. To evaluate statistical significance, we performed the Wilcoxon signed rank test [13], which confirmed significant reductions in average variance ( $p = 0.002$ ),  $\text{CHAIR}_s$  ( $p = 0.002$ ), and  $\text{CHAIR}_i$  ( $p = 0.004$ ), all statistically significant at  $p < 0.05$ . The F1 score was preserved. These results demonstrate that the uncertainty of visual tokens contributes to object hallucination, and that our method effectively suppresses this uncertainty, thereby mitigating hallucinations in LVLMs.

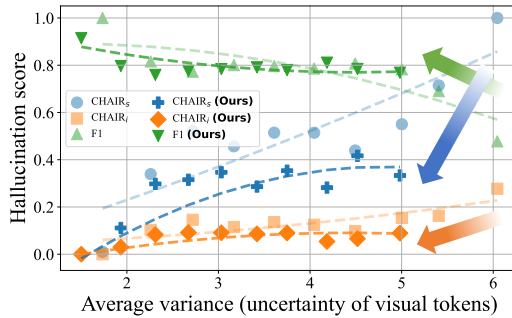


Figure 6: **Impact of the proposed masking strategy on visual token uncertainty.** Average token-level variance estimated via MC dropout decreases after applying our method, indicating reduced uncertainty. This reduction correlates with improved performance on object hallucination metrics. The trend line was fitted with quadratic function.

Table 1: **Quantitative results on CHAIR and POPE benchmarks.** Object hallucination is evaluated on the CHAIR and POPE benchmarks using three LVLMs and five decoding strategies, both with and without our method. POPE results are reported on three splits: Random, Popular, and Adversarial. The maximum token length is set to 512.  $\Delta\%$  denotes the relative difference in performance.  $\uparrow / \downarrow$  indicate that higher/lower values are better. We highlight the best scores in **bold**.

| Method       | Greedy           |             |             | OPERA               |             |             | VCD                 |       |            | PAI                 |             |             | Devils              |       |             |                     |
|--------------|------------------|-------------|-------------|---------------------|-------------|-------------|---------------------|-------|------------|---------------------|-------------|-------------|---------------------|-------|-------------|---------------------|
|              | Orig.            | +Ours       | $\Delta\%$  | Orig.               | +Ours       | $\Delta\%$  | Orig.               | +Ours | $\Delta\%$ | Orig.               | +Ours       | $\Delta\%$  | Orig.               | +Ours | $\Delta\%$  |                     |
| LLaVA-1.5-7B | $C_s \downarrow$ | 47.4        | 29.2        | $\downarrow 38.4\%$ | 47.8        | 29.4        | $\downarrow 38.5\%$ | 53.8  | 35.2       | $\downarrow 34.6\%$ | 33.2        | 26.0        | $\downarrow 21.7\%$ | 27.0  | <b>23.0</b> | $\downarrow 14.8\%$ |
|              | $C_i \downarrow$ | 12.2        | 9.3         | $\downarrow 23.8\%$ | 12.8        | 9.5         | $\downarrow 25.8\%$ | 15.2  | 10.7       | $\downarrow 29.6\%$ | 8.5         | 7.9         | $\downarrow 7.1\%$  | 6.6   | <b>5.6</b>  | $\downarrow 15.2\%$ |
|              | F1 $\uparrow$    | 77.9        | 78.2        | $\uparrow 0.4\%$    | 77.7        | <b>78.4</b> | $\uparrow 0.9\%$    | 75.2  | 75.2       | $\uparrow 0.0\%$    | 78.3        | 77.2        | $\downarrow 1.4\%$  | 78.3  | 78.0        | $\downarrow 0.4\%$  |
|              | Rand. $\uparrow$ | 89.3        | 89.3        | $\uparrow 0.0\%$    | 89.2        | 88.6        | $\downarrow 0.7\%$  | 84.6  | 86.2       | $\uparrow 1.9\%$    | 89.4        | 89.2        | $\downarrow 0.2\%$  | 89.6  | <b>90.0</b> | $\uparrow 0.4\%$    |
|              | Pop. $\uparrow$  | 85.8        | 85.8        | $\uparrow 0.0\%$    | 85.8        | 85.2        | $\downarrow 0.7\%$  | 82.4  | 82.9       | $\uparrow 0.6\%$    | 86.0        | 86.4        | $\uparrow 0.5\%$    | 86.4  | <b>87.2</b> | $\uparrow 0.9\%$    |
|              | Adv. $\uparrow$  | 79.3        | 80.0        | $\uparrow 0.9\%$    | <b>80.3</b> | 79.6        | $\downarrow 0.9\%$  | 77.0  | 78.1       | $\uparrow 1.4\%$    | 79.5        | 79.9        | $\uparrow 0.5\%$    | 78.6  | 79.6        | $\uparrow 1.3\%$    |
| Shikra-7B    | $C_s \downarrow$ | 58.0        | 43.2        | $\downarrow 25.5\%$ | 34.8        | 28.8        | $\downarrow 17.2\%$ | 56.2  | 47.2       | $\downarrow 16.0\%$ | 32.4        | 22.2        | $\downarrow 31.5\%$ | 24.4  | <b>20.6</b> | $\downarrow 15.6\%$ |
|              | $C_i \downarrow$ | 15.6        | 11.7        | $\downarrow 25.0\%$ | 11.1        | 9.6         | $\downarrow 13.5\%$ | 16.1  | 12.8       | $\downarrow 20.5\%$ | 7.8         | <b>6.1</b>  | $\downarrow 21.8\%$ | 7.6   | 6.8         | $\downarrow 10.5\%$ |
|              | F1 $\uparrow$    | 74.7        | <b>76.9</b> | $\uparrow 2.9\%$    | 74.2        | 74.2        | $\uparrow 0.0\%$    | 74.4  | 75.2       | $\uparrow 1.1\%$    | 76.7        | 75.1        | $\downarrow 2.1\%$  | 73.3  | 72.2        | $\downarrow 1.5\%$  |
|              | Rand. $\uparrow$ | 83.2        | 85.1        | $\uparrow 2.3\%$    | 84.8        | <b>85.4</b> | $\uparrow 0.7\%$    | 82.1  | 82.7       | $\uparrow 0.7\%$    | 83.9        | 84.0        | $\uparrow 0.1\%$    | 83.8  | 82.5        | $\downarrow 1.6\%$  |
|              | Pop. $\uparrow$  | 82.3        | 82.6        | $\uparrow 0.4\%$    | 82.8        | 82.1        | $\downarrow 0.8\%$  | 79.7  | 80.7       | $\uparrow 1.3\%$    | <b>83.1</b> | 80.7        | $\downarrow 2.9\%$  | 79.9  | 78.2        | $\downarrow 2.1\%$  |
|              | Adv. $\uparrow$  | 78.2        | 78.8        | $\uparrow 0.8\%$    | 79.2        | <b>79.7</b> | $\uparrow 0.6\%$    | 77.3  | 77.1       | $\downarrow 0.3\%$  | 78.8        | 77.4        | $\downarrow 1.8\%$  | 77.7  | 76.7        | $\downarrow 1.3\%$  |
| MiniGPT-4    | $C_s \downarrow$ | 28.6        | 27.4        | $\downarrow 4.2\%$  | 23.8        | 22.6        | $\downarrow 5.0\%$  | 32.0  | 30.6       | $\downarrow 4.4\%$  | 19.6        | <b>17.8</b> | $\downarrow 9.2\%$  | 21.6  | 20.8        | $\downarrow 3.7\%$  |
|              | $C_i \downarrow$ | 8.5         | 8.3         | $\downarrow 2.4\%$  | 8.8         | 8.5         | $\downarrow 3.4\%$  | 9.7   | 9.1        | $\downarrow 6.2\%$  | 6.2         | <b>6.0</b>  | $\downarrow 3.2\%$  | 7.5   | 7.0         | $\downarrow 6.7\%$  |
|              | F1 $\uparrow$    | 71.5        | 71.3        | $\downarrow 0.3\%$  | 69.8        | 70.0        | $\uparrow 0.3\%$    | 70.2  | 71.3       | $\uparrow 1.7\%$    | <b>71.7</b> | <b>71.7</b> | $\uparrow 0.0\%$    | 70.1  | 70.4        | $\uparrow 0.4\%$    |
|              | Rand. $\uparrow$ | <b>82.8</b> | 82.5        | $\downarrow 0.4\%$  | 74.2        | 74.4        | $\uparrow 0.3\%$    | 59.2  | 59.3       | $\uparrow 0.2\%$    | 82.1        | 82.0        | $\downarrow 0.1\%$  | 77.4  | 77.8        | $\uparrow 0.5\%$    |
|              | Pop. $\uparrow$  | 75.1        | 74.6        | $\downarrow 0.7\%$  | 71.3        | 71.8        | $\uparrow 0.7\%$    | 54.9  | 55.0       | $\uparrow 0.2\%$    | <b>75.8</b> | 75.2        | $\downarrow 0.8\%$  | 68.4  | 68.6        | $\uparrow 0.3\%$    |
|              | Adv. $\uparrow$  | 71.8        | 71.2        | $\downarrow 0.8\%$  | 69.7        | 69.4        | $\downarrow 0.4\%$  | 53.8  | 54.2       | $\uparrow 1.1\%$    | <b>72.1</b> | 71.6        | $\downarrow 0.7\%$  | 65.2  | 65.3        | $\uparrow 0.2\%$    |

## 4 Experiments

### 4.1 Experimental Setup

**Baselines and implementation details.** We evaluate our method on diverse LVLMs differing in size, architecture, and vision encoders: LLaVA-1.5-7B [35] with CLIP-L/336px [46], Shikra-7B [9] with CLIP-L, and MiniGPT-4 using EVACLIP-g [51] and a Q-Former for image-text alignment. To assess compatibility, we integrate our method with hallucination mitigation methods including OPERA[20], VCD [27], PAI [37], and Devils [24]. Adversarial attacks are run with  $k = 3$  and 200 PGD steps. The uncertainty masks  $M$  are extracted from layers  $\mathcal{S} = \{1, \dots, 10\}$  of the vision encoder, with masking applied to layers 13–17 for LLaVA-1.5 and Shikra, and 9–16 for MiniGPT-4.  $\sigma_{\text{th}s}$  are tuned per baseline-method pair. Further details are provided in Appendix Sec. C, and D.3.

**Benchmarks.** To measure object hallucination, we use three standard benchmarks. CHAIR [47] measures sentence-level ( $C_s := \text{CHAIR}_s$ ) and instance-level ( $C_i := \text{CHAIR}_i$ ) hallucinations from generated descriptions with 500 prompts randomly sampled from COCO [33]:

$$\text{CHAIR}_s = \frac{|\{\text{sentences with hallucinated objects}\}|}{|\{\text{all sentences}\}|}, \quad \text{CHAIR}_i = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all mentioned objects}\}|}. \quad (5)$$

POPE [31] evaluates hallucination through binary object presence queries across three splits (Random, Popular, Adversarial), total 9,000 prompts, reporting accuracy. AMBER [56] comprehensively evaluates hallucination in two settings: a generative approach (Gen.) that assesses hallucination through image captioning and a discriminative approach (Disc.) that uses yes/no choices. To measure object hallucination with AMBER, we adopted the full set of Gen. and the ‘Existence’ subset of Disc., conducting with a total of 5,928 prompts. See Appendix D.1 and G.2 for more details and results.

### 4.2 Experimental results

**Quantitative Results.** We evaluate the effectiveness of our method in mitigating object hallucinations in multiple LVLM using CHAIR [47] and POPE [31] benchmarks. As shown in Table 1, our method consistently reduces hallucination rates ( $C_s$ ,  $C_i$ ) across LLaVA-1.5-7B, Shikra-7B, and MiniGPT-4, while preserving caption quality (F1). For example, on LLaVA-1.5-7B,  $C_s$  drops from 47.4 to 29.2 and  $C_i$  from 12.2 to 9.3. Although the improvement on MiniGPT-4 is smaller, this is

Table 2: **Quantitative results on AMBER benchmark for LLaVA-1.5-7B.** We evaluate object hallucination using the AMBER benchmark under various mitigation methods, including combinations with our approach. AMBER measures hallucination in generative (Gen.) and discriminative (Disc.) settings, with its score offering a comprehensive assessment across both. The maximum token length is set to 512 for generative task.  $\Delta\%$  denotes the relative difference in performance.

| Method |         | Greedy |       |            | OPERA |             |            | VCD   |       |            | PAI   |       |            | Devils     |             |            |
|--------|---------|--------|-------|------------|-------|-------------|------------|-------|-------|------------|-------|-------|------------|------------|-------------|------------|
|        |         | Orig.  | +Ours | $\Delta\%$ | Orig. | +Ours       | $\Delta\%$ | Orig. | +Ours | $\Delta\%$ | Orig. | +Ours | $\Delta\%$ | Orig.      | +Ours       | $\Delta\%$ |
| Gen.   | CHAIR ↓ | 6.7    | 5.1   | ↓23.9%     | 7.4   | 5.8         | ↓21.6%     | 8.5   | 6.1   | ↓28.2%     | 5.1   | 4.7   | ↓7.8%      | 4.1        | <b>3.9</b>  | ↓4.9%      |
|        | Hal ↓   | 30.2   | 24.2  | ↓19.9%     | 33.0  | 23.3        | ↓29.4%     | 38.4  | 28.6  | ↓25.5%     | 25.1  | 22.5  | ↓10.4%     | 21.0       | <b>20.9</b> | ↓0.5%      |
|        | Cog ↓   | 3.8    | 2.3   | ↓39.5%     | 3.7   | 2.1         | ↓43.2%     | 4.4   | 2.3   | ↓47.7%     | 1.9   | 1.9   | ↓0.0%      | <b>1.4</b> | 1.5         | ↑7.1%      |
| Disc.  | Pre. ↑  | 100.0  | 100.0 | ↑0.0%      | 100.0 | 100.0       | ↑0.0%      | 100.0 | 100.0 | ↑0.0%      | 100.0 | 100.0 | ↑0.0%      | 100.0      | 100.0       | ↑0.0%      |
|        | Rec. ↑  | 71.2   | 78.0  | ↑9.6%      | 74.9  | <b>81.0</b> | ↑7.5%      | 67.3  | 75.7  | ↑12.5%     | 71.9  | 74.1  | ↑3.1%      | 72.5       | 75.2        | ↑3.7%      |
|        | F1 ↑    | 83.2   | 87.6  | ↑5.3%      | 85.6  | <b>89.5</b> | ↑4.6%      | 80.4  | 86.2  | ↑7.2%      | 83.6  | 85.1  | ↑1.8%      | 84.1       | 85.8        | ↑2.0%      |
| AMBER↑ |         | 88.2   | 91.2  | ↑3.4%      | 89.1  | <b>91.8</b> | ↑3.0%      | 86.0  | 90.1  | ↑4.8%      | 89.2  | 90.2  | ↑1.1%      | 90.0       | 91.0        | ↑1.1%      |

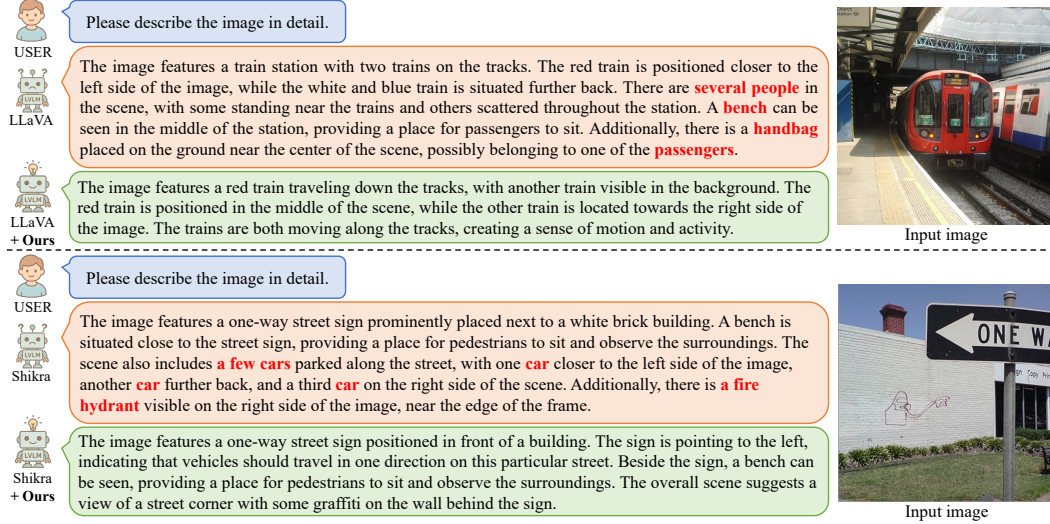


Figure 7: **Qualitative results of our method on LLaVA-1.5-7B and Shikra-7B.** Greedy decoding leads to object hallucinations by describing non-existent objects in the image (e.g., ‘several people’, ‘bench’, ‘handbag’, ‘passengers’ in LLaVA; ‘a few cars’, ‘car’, ‘a fire hydrant’ in Shikra). In contrast, our method, which modifies only the vision encoder, substantially reduces such hallucinations.

likely due to its Q-Former module between the vision encoder and LLM, which limits the effect of our method modifying the vision encoder. In POPE, our method yields comparable or slightly improved performance across all models, indicating robustness under discriminative evaluation settings. Furthermore, it integrates well with existing mitigation methods such as OPERA, VCD, PAI, and Devils, providing additional gains without compromising caption quality. We also present results on newer models (DeepSeek-VL [38], Qwen2.5-VL [4]), and larger models (LLaVA-1.5-13B) are provided in Appendix Sec. G.2.

We further evaluate our method on the AMBER benchmark [56] using LLaVA-1.5-7B across five strategies as depicted in Table 2. Our approach substantially reduces object hallucinations in both generative and discriminative tasks, achieving up to a 28.2% reduction in CHAIR and a 7.2% improvement in F1, resulting in consistently higher AMBER scores across all settings.

**Qualitative results.** We provide qualitative examples to demonstrate the effectiveness of our method. As shown in Fig. 7, greedy decoding with vanilla LVLMs leads to object hallucinations, generating descriptions that mention non-existent objects such as *several people*, *bench*, *car*, or *a fire hydrant*. In contrast, our method substantially reduces such hallucinations in the generated outputs. Notably, in the case of Shikra integrated with our method, the model is able to correctly identify previously overlooked objects like *graffiti*, reflecting improved visual grounding and descriptiveness. We provide further qualitative results for various combinations of models and methods in the Appendix Sec. G.3.

Table 3: **Impact of vision encoder layers on generating the uncertainty mask  $M$ .** Using early layers of vision encoder (1–10) to compute  $M$  yields the most effective object hallucination mitigation performance.

| Mask Source Layer  | $C_s \downarrow$ | $C_i \downarrow$ | F1 $\uparrow$ |
|--------------------|------------------|------------------|---------------|
| Greedy             | 47.4             | 12.2             | 77.9          |
| <b>Layers 1–10</b> | <b>29.2</b>      | <b>9.3</b>       | <b>78.2</b>   |
| Layers 11–20       | 44.2             | 12.7             | 77.4          |
| Layers 21–22       | 41.8             | 12.1             | 77.9          |

Table 4: **Effect of applying the uncertainty mask  $M$  to different layers in the vision encoder.** Applying the mask at middle layers of vision encoder (13–17) results in the most effective performance.

| Masking Layer Range | $C_s \downarrow$ | $C_i \downarrow$ | F1 $\uparrow$ |
|---------------------|------------------|------------------|---------------|
| Greedy              | 47.4             | 12.2             | 77.9          |
| Layers 1–8          | 45.0             | 12.6             | 77.9          |
| Layers 8–12         | 55.8             | 15.5             | 75.7          |
| <b>Layers 13–17</b> | <b>29.2</b>      | <b>9.3</b>       | <b>78.2</b>   |
| Layers 18–22        | 45.8             | 13.0             | 77.7          |

Table 5: **Comparison of masking strategies for uncertain visual tokens.** We compare our attention-level masking method with alternatives applied at different stages of the vision encoder (VE). S.M. denotes soft masking, which attenuates uncertain tokens by a small factor (e.g., 0.1 or 0.2).

| Strategy         | Greedy | Input of VE | Output of VE | MLP Layer | S.M. (0.1 / 0.2)   | <b>Ours</b> |
|------------------|--------|-------------|--------------|-----------|--------------------|-------------|
| $C_s \downarrow$ | 47.4   | 47.4        | 34.4         | 51.0      | 35.0 / 40.0        | <b>29.2</b> |
| $C_i \downarrow$ | 12.2   | 12.5        | 10.0         | 13.5      | 10.4 / 11.5        | <b>9.3</b>  |
| F1 $\uparrow$    | 77.9   | 77.5        | 74.7         | 77.9      | <b>78.3</b> / 78.1 | <b>78.2</b> |

### 4.3 Ablation Study and Analysis

To assess the impact of each component on reducing object hallucination, we perform ablation studies on the LLaVA-1.5-7B [35] model. We examine two key factors in the vision encoder: (1) uncertain visual token estimation and (2) a training-free masking strategy. Each experiment isolates one variable to ensure fair comparison. Limitations of our method are discussed in Appendix J.

**Uncertainty estimation of visual tokens from early layers of vision encoder.** We examine which layers of vision encoder are most effective for generating the binary uncertainty mask  $M$  using PGD-based adversarial attacks. As shown in Table 3, extracting uncertainty from early layers (1 to 10) leads to the largest reduction in hallucinations ( $C_s$ ,  $C_i$ ) and the highest F1 score, outperforming intermediate or deeper layers. This result aligns with Sec.3.1.2 and Fig.3, where early layers exhibit smaller adversarial feature shifts, making them more suitable for uncertainty estimation.

**Masking uncertain visual tokens in intermediate layers of vision encoder.** We investigate the effect of applying the binary uncertainty mask  $M$  to different layers of self-attention process within the vision encoder. As shown in Table 4, masking at intermediate layers (13 to 17) yields the best performance, significantly reducing hallucination ( $C_s$ ,  $C_i$ ) and achieving the highest F1 score. In contrast, masking in earlier layers shows limited benefit, and deeper layers provide minimal gains.

**Comparative analysis of masking strategies for uncertain visual tokens.** We compare several masking strategies using the binary uncertainty mask  $M$ , including masking at the input image, the output of the vision encoder, the MLP layer before the residual connection in the transformer block, and soft masking applied to the self-attention that attenuates uncertain visual tokens by a small factor. As shown in Table 5, our method, which applies hard masking within the self-attention mechanism using  $M$ , achieves the best hallucination scores while maintaining a competitive F1 score.

## 5 Conclusion

We present a simple yet effective approach for mitigating object hallucination in Large Vision-Language Models (LVLMs) by identifying uncertain visual tokens within the vision encoder and reducing their influence through masking these tokens in their self-attention layers. Our theoretical and empirical analyses show that adversarial perturbations efficiently approximate an upper bound of epistemic uncertainty, which we confirm to be strongly correlated with object hallucination in LVLMs. Extensive experiments demonstrate that our approach consistently reduces object hallucination across diverse models and integrates seamlessly with other prior arts to improve performance.



## Acknowledgement

This work was supported in part by Institute of Information & communications Technology Planning & Evaluation (IITP) grants funded by the Korea government(MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)], (No.RS-2025-02314125, Effective Human-Machine Teaming With Multimodal Hazy Oracle Models) and by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2025-02263628). Also, the authors acknowledged the financial support from the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University.

## References

- [1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *ICCV*, pages 8948–8957, 2019.
- [2] Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Guang Dai, Ping Chen, and Shijian Lu. Agla: Mitigating object hallucinations in large vision-language models with assembly of global and local attention. *arXiv preprint arXiv:2406.12718*, 2024.
- [3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.
- [4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibong Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [5] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.
- [6] Zikui Cai, Yaoteng Tan, and M Salman Asif. Ensemble-based blackbox attacks on dense prediction. In *CVPR*, pages 4045–4055, 2023.
- [7] Nicholas Carlini, Milad Nasr, Christopher A Choquette-Choo, Matthew Jagielski, Irena Gao, Pang Wei W Koh, Daphne Ippolito, Florian Tramèr, and Ludwig Schmidt. Are aligned neural networks adversarially aligned? *NeurIPS*, 36:61478–61500, 2023.
- [8] Liwei Che, Tony Qingze Liu, Jing Jia, Weiye Qin, Ruixiang Tang, and Vladimir Pavlovic. Eazy: Eliminating hallucinations in llms by zeroing out hallucinatory image tokens. *arXiv preprint arXiv:2503.07772*, 2025.
- [9] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm’s referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.
- [10] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. In *ICLR*, 2023.
- [11] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality, March 2023.
- [12] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In *ICLR*, 2024.
- [13] William Jay Conover. *Practical nonparametric statistics*. John Wiley & sons, 1999.
- [14] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023.
- [15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.



- [16] Jinhao Duan, Fei Kong, Hao Cheng, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. Truthprint: Mitigating lvlm object hallucination via latent truthful-guided pre-intervention. *arXiv preprint arXiv:2503.10602*, 2025.
- [17] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *CVPR*, pages 19358–19369, 2023.
- [18] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *ICLR*, 2015.
- [19] Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *AAAI*, volume 38, pages 18135–18143, 2024.
- [20] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *CVPR*, pages 13418–13427, 2024.
- [21] Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-introspective decoding: Alleviating hallucinations for large vision-language models. *arXiv preprint arXiv:2408.02032*, 2024.
- [22] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. If you use this software, please cite it as below.
- [23] Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaying Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. In *CVPR*, pages 27036–27046, 2024.
- [24] Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. *arXiv preprint arXiv:2411.16724*, 2024.
- [25] Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *ICLR*, 2025.
- [26] Max-Heinrich Laves, Sontje Ihler, Karl-Philipp Kortmann, and Tobias Ortmaier. Calibration of model uncertainty for dropout variational inference. *arXiv preprint arXiv:2006.11584*, 2020.
- [27] 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. In *CVPR*, pages 13872–13882, 2024.
- [28] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.
- [29] Jiaming Li, Jiacheng Zhang, Zequn Jie, Lin Ma, and Guanbin Li. Mitigating hallucination for large vision language model by inter-modality correlation calibration decoding. *arXiv preprint arXiv:2501.01926*, 2025.
- [30] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pages 19730–19742. PMLR, 2023.
- [31] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, 2023.
- [32] Siyuan Liang, Baoyuan Wu, Yanbo Fan, Xingxing Wei, and Xiaochun Cao. Parallel rectangle flip attack: A query-based black-box attack against object detection. In *ICCV*, pages 7677–7687. IEEE Computer Society, 2021.
- [33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.
- [34] Hanchao Liu, Wenyan 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. *arXiv preprint arXiv:2402.00253*, 2024.

- [35] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pages 26296–26306, 2024.
- [36] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36:34892–34916, 2023.
- [37] Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in vlms. In *ECCV*, pages 125–140. Springer, 2024.
- [38] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.
- [39] Xinyu Lyu, Beita Chen, Lianli Gao, Hengtao Shen, and Jingkuan Song. Alleviating hallucinations in large vision-language models through hallucination-induced optimization. *NeurIPS*, 37:122811–122832, 2024.
- [40] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.
- [41] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.
- [42] Shunqi Mao, Chaoyi Zhang, and Weidong Cai. Through the magnifying glass: Adaptive perception magnification for hallucination-free vlm decoding. *arXiv preprint arXiv:2503.10183*, 2025.
- [43] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, pages 3195–3204, 2019.
- [44] Jishnu Mukhoti and Yarin Gal. Evaluating bayesian deep learning methods for semantic segmentation. *arXiv preprint arXiv:1811.12709*, 2018.
- [45] Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models. In *AAAI*, volume 38, pages 21527–21536, 2024.
- [46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021.
- [47] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *EMNLP*, pages 4035–4045, 2018.
- [48] Alfred P Rovai, Jason D Baker, and Michael K Ponton. *Social science research design and statistics: A practitioner’s guide to research methods and IBM SPSS*. Watertree Press LLC, 2013.
- [49] Christian Schlarmann and Matthias Hein. On the adversarial robustness of multi-modal foundation models. In *ICCV*, pages 3677–3685, 2023.
- [50] Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. In *ICLR*, 2024.
- [51] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.
- [52] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.
- [53] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.
- [54] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shriti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- [55] Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. Mllm can see? dynamic correction decoding for hallucination mitigation. *arXiv preprint arXiv:2410.11779*, 2024.
- [56] Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *CoRR*, 2023.

- [57] Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*, 2023.
- [58] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *CVPR*, pages 19175–19186, 2023.
- [59] Yubo Wang, Chaohu Liu, Yanqiu Qu, Haoyu Cao, Deqiang Jiang, and Linli Xu. Break the visual perception: Adversarial attacks targeting encoded visual tokens of large vision-language models. In *ACM MM*, pages 1072–1081, 2024.
- [60] Chunzhao Xie, Tongxuan Liu, Lei Jiang, Yuting Zeng, Yunheng Shen, Weizhe Huang, Jing Li, Xiaohua Xu, et al. Tarac: Mitigating hallucination in vlms via temporal attention real-time accumulative connection. *arXiv preprint arXiv:2504.04099*, 2025.
- [61] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *ICCV*, pages 1369–1378, 2017.
- [62] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yeqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.
- [63] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.
- [64] Zihao Yue, Liang Zhang, and Qin Jin. Less is more: Mitigating multimodal hallucination from an eos decision perspective. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11766–11781, 2024.
- [65] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pages 11975–11986, 2023.
- [66] Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. Mllms know where to look: Training-free perception of small visual details with multimodal llms. *arXiv preprint arXiv:2502.17422*, 2025.
- [67] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. *NeurIPS*, 36:54111–54138, 2023.
- [68] 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. In *ICLR*, 2024.
- [69] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *ICLR*, 2024.
- [70] Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding. *arXiv preprint arXiv:2402.18476*, 2024.

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: We introduced our approach to mitigate object hallucination in title, abstract, and introduction. Also, we summarized our contributions explicitly in Sec. 1.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: We discussed limitation of our work in Appendix Sec J.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: We provide the complete set of assumptions and full proofs for Lemma 3.1 and Theorem 3.2, with appropriate references to the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: We provide code for reproducibility and detailed implementation details in Appendix Sec. C, benchmarks and baseline models in Appendix Sec. D, along with the corresponding GitHub link.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

#### 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide our code implementation in supplementary materials.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide experimental setting and details in Appendix Sec. C and Sec. D.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We provide error bars on our experimental results in Fig. 3, and report our results’ statistical significance of Fig. 4 in Sec. 3.1.2 and Sec. 3.2.3.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer “Yes” if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).



- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

#### 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We specify in Appendix Sec. C that all main experiments were conducted using an NVIDIA A100 GPU with 80GB of memory.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

#### 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: We comply with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

#### 10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We discuss both the potential positive and negative societal impacts of our work in Appendix Sec. I.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our paper does not involve releasing any models or datasets that pose a high risk of misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We properly cite all utilized code, benchmark datasets, and models, and provide the corresponding GitHub links in Appendix Sec. D.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

### 13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [\[Yes\]](#)

Justification: We will provide a README file alongside the released code in the supplementary materials, which includes usage instructions, details of the benchmark datasets, and descriptions of the models used in our experiments.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

### 14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [\[NA\]](#)

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

### 15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [\[NA\]](#)

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

#### 16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: Our core method development in this research does not incorporate large language models as any essential, novel, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

## A Proofs

### A.1 Proof of Lemma 3.1

**Lemma 3.1** (Approximate local Gaussianity under small perturbation). *Let  $f = \{f_t\}_{t=1}^L$  be a smooth  $L$ -layer neural network parameterized by  $\theta$ . For an input  $x \in \mathbb{R}^{N \times 3}$ , define the hidden state at layer  $t$  as  $z^{(t)} = f_t \circ \dots \circ f_1(x)$ . For a perturbed input  $x + \epsilon$ , with  $\|\epsilon\|_\infty \leq k$  for sufficiently small  $k > 0$ , define the perturbed hidden state as  $Z^{(t)} = f_t \circ \dots \circ f_1(x + \epsilon)$ . Then, under the assumption that the perturbation is small and  $f \in C^2$ ,  $Z^{(t)}$  can be locally approximated by a Gaussian centered at  $z^{(t)}$ , with a third-order remainder in the log-density.*

*Proof.* Let  $f = \{f_t\}_{t=1}^L$  be a smooth  $L$ -layer neural network parameterized by  $\theta$ , and let  $z^{(t)} := f_t \circ \dots \circ f_1(x)$  denote the hidden state at layer  $t$  for a clean input  $x \in \mathbb{R}^{N \times 3}$ . For a perturbed input  $x + \epsilon$ , with  $\|\epsilon\|_\infty \leq k$  for small  $k > 0$ , define the perturbed hidden state as  $Z^{(t)} := f_t \circ \dots \circ f_1(x + \epsilon)$ .

For the clean and perturbed inputs, define

$$y^* := f(x; \theta) = f^{(t)}(z^{(t)}; \theta^{(t)}), \quad y := f(x + \epsilon; \theta) = f^{(t)}(z^{(t)} + \epsilon'; \theta^{(t)}), \quad (\text{A1})$$

where  $f^{(t)} = f_L \circ \dots \circ f_{t+1}$ ,  $\theta^{(t)}$  are its parameters, and  $\epsilon'$  is the residual vector at layer  $t$  induced by the input perturbation  $\epsilon$ . The perturbation  $\epsilon$  is chosen to maximize the adversarial objective  $C\|y - y^*\|_2^2$ , or equivalently minimize  $\exp(-C\|y - y^*\|_2^2)$ , under  $\|\epsilon\|_\infty \leq k$ .

Motivated by this, we approximate the conditional distribution of hidden states near  $z^{(t)}$  using a local energy-based form,

$$p_\theta(z | y^*) \propto \exp(-C\|f^{(t)}(z; \theta^{(t)}) - f^{(t)}(z^{(t)}; \theta^{(t)})\|_2^2), \quad (\text{A2})$$

for  $z$  in a neighborhood of  $z^{(t)}$ . Since  $f$  is twice continuously differentiable, the conditional log-density admits a second-order Taylor expansion around  $z^{(t)}$ :

$$\begin{aligned} \log p_\theta(z | y^*) &= \log p_\theta(z^{(t)} | y^*) + (z - z^{(t)})^\top \nabla_z \log p_\theta(z | y^*)|_{z=z^{(t)}} \\ &\quad + \frac{1}{2}(z - z^{(t)})^\top H^{(t)}(z - z^{(t)}) + R(z), \end{aligned} \quad (\text{A3})$$

where  $H^{(t)} := \nabla_z^2 \log p_\theta(z | y^*)|_{z=z^{(t)}}$  is the Hessian and  $R(z) = \mathcal{O}(\|z - z^{(t)}\|^3)$ .

The first-order term vanishes as follows:

$$\nabla_z \log p_\theta(z | y^*)|_{z=z^{(t)}} = -2C \cdot J_{f^{(t)}}(z; \theta^{(t)})^\top \left( f^{(t)}(z; \theta^{(t)}) - f^{(t)}(z^{(t)}; \theta^{(t)}) \right)|_{z=z^{(t)}} = 0. \quad (\text{A4})$$

Therefore,

$$\log p_\theta(z | y^*) = \log p_\theta(z^{(t)} | y^*) + \frac{1}{2}(z - z^{(t)})^\top H^{(t)}(z - z^{(t)}) + R(z). \quad (\text{A5})$$

The quadratic term coincides with the log-density of a Gaussian centered at  $z^{(t)}$  with covariance  $(-H^{(t)})^{-1}$ , while the remainder  $R(z)$  is of order  $\mathcal{O}(\|z - z^{(t)}\|^3)$ .

Therefore, the perturbed hidden state  $Z^{(t)}$  under small input perturbations can be locally approximated by a Gaussian centered at  $z^{(t)}$ , with approximation error of third order in the log-density.  $\square$

### A.2 Proof of Theorem 3.2

**Theorem 3.2** (Upper bound of differential entropy increases as hidden state deviation increases under adversarial attack). *Let  $x$  be an input image, and let  $\epsilon$  be a small adversarial perturbation. Define the perturbed input as  $X := x + \epsilon$ . Let  $f = \{f_t\}_{t=1}^L$  be a smooth  $L$ -block transformer that processes a sequence of  $N$  input tokens. Let  $z^{(t)} := f_t \circ \dots \circ f_1(x) \in \mathbb{R}^{N \times d}$  and  $Z^{(t)} := f_t \circ \dots \circ f_1(X) \in \mathbb{R}^{N \times d}$  be the hidden states at layer  $t$  for the clean and perturbed inputs, respectively. Denote the  $i$ -th token representation at layer  $t$  as  $z_i^{(t)} \in \mathbb{R}^d$  and  $Z_i^{(t)} \in \mathbb{R}^d$ . If  $Z_i^{(t)}$  changes smoothly with small  $\epsilon$ , then the upper bound of the differential entropy of  $Z_i^{(t)}$  increases as  $\mathbb{E}_\epsilon[\|Z_i^{(t)} - z_i^{(t)}\|_2^2]$  increases.*

*Proof.* Let  $x$  be an input image and  $\epsilon$  a small perturbation satisfying  $\|\epsilon\|_\infty \leq k$ , where  $k$  is sufficiently small for a first-order Taylor expansion. Define

$$z_i^{(t)} := f_i^{(t)}(x), \quad Z_i^{(t)} := f_i^{(t)}(x + \epsilon), \quad (\text{A6})$$

where  $f_i^{(t)}$  denotes the hidden state of token  $i$  at layer  $t$ , and  $f = f_t \circ \dots \circ f_1$  is assumed to be twice continuously differentiable.

By the multivariate Taylor expansion of  $f_i^{(t)}(x + \epsilon)$  around  $x$ , we have

$$Z_i^{(t)} = z_i^{(t)} + J_i^{(t)}\epsilon + R_i^{(t)}(\epsilon), \quad (\text{A7})$$

where  $J_i^{(t)} := \left. \frac{\partial z_i^{(t)}}{\partial x} \right|_x \in \mathbb{R}^{d \times D}$  is the Jacobian matrix, and  $\|R_i^{(t)}(\epsilon)\| = \mathcal{O}(\|\epsilon\|^2)$ .

With the assumption of the perturbation upper bound  $k$ , the remainder  $R_i^{(t)}(\epsilon)$  is negligible compared to the linear term. Under this assumption, we define the deviation:

$$\Delta Z_i^{(t)} := Z_i^{(t)} - z_i^{(t)} = J_i^{(t)}\epsilon. \quad (\text{A8})$$

Let  $\Sigma_\epsilon := \mathbb{E}[\epsilon\epsilon^\top]$ . Then the covariance of  $\Delta Z_i^{(t)}$  is

$$\Sigma_{\Delta Z_i^{(t)}} := \text{Cov}[\Delta Z_i^{(t)}] = J_i^{(t)}\Sigma_\epsilon(J_i^{(t)})^\top. \quad (\text{A9})$$

By the local Gaussianity assumption (Lemma 3.1),  $Z_i^{(t)}$  can be approximated as a multivariate Gaussian. Hence, by the entropy formula for multivariate Gaussians, the differential entropy is

$$h(Z_i^{(t)}) = \frac{1}{2} \log \left( (2\pi e)^d \cdot \det(\Sigma_{\Delta Z_i^{(t)}}) \right). \quad (\text{A10})$$

Applying the AM–GM inequality to the eigenvalues of  $\Sigma_{\Delta Z_i^{(t)}}$ , we obtain

$$\det(\Sigma_{\Delta Z_i^{(t)}})^{1/d} \leq \frac{1}{d} \text{tr}(\Sigma_{\Delta Z_i^{(t)}}) = \frac{1}{d} \mathbb{E}[\|\Delta Z_i^{(t)}\|_2^2]. \quad (\text{A11})$$

Thus, the entropy is bounded as:

$$h(Z_i^{(t)}) \leq \frac{d}{2} \log \left( \frac{1}{d} \mathbb{E}[\|\Delta Z_i^{(t)}\|_2^2] \right) + C, \quad (\text{A12})$$

where  $C = \frac{d}{2} \log(2\pi e)$  is a constant.

Hence, the upper bound of the entropy increases as  $\mathbb{E}[\|\Delta Z_i^{(t)}\|_2^2]$  increases, which completes the proof.  $\square$

### A.3 On practicality of the proved upper bound

Assuming that the deviation of hidden states follows a Gaussian distribution, the differential entropy of each token is proportional to the determinant of the covariance matrix  $\Sigma_{\Delta Z}$ . However, our empirical analysis reveals that this covariance matrix is highly low-rank. By decomposing the covariance matrix obtained from 2048 adversarial attacks on the visual tokens of 100 images with LLaVA-1.5-7B [28] using PCA, we found that the top 8 components ( $8/1024 = 0.8\%$  of the total dimension) account for  $94.2\%$  ( $\pm 0.4\%$ ) of the total variance, with most eigenvalues close to zero. Under such conditions, computing  $\det(\Sigma_{\Delta Z})$  for entropy estimation becomes numerically unstable, as values underflow to zero, making direct entropy comparison infeasible. In contrast, using  $\text{tr}(\Sigma_{\Delta Z})$  provides a numerically stable alternative that is theoretically well-grounded under anisotropy and preserves token-wise uncertainty ordering. This trace-based measure also aligns with the qualitative uncertainty maps in Fig. 2, further supporting its practical validity.

## B Code

To support reproducibility, we include the implementation of our method in the supplementary material. Detailed instructions for running the code and setting up the environment are provided in the accompanying `README.md` file.

## C Implementation Details

As our method is designed to work in conjunction with various LVLMs and existing mitigation methods such as OPERA, VCD, PAI and Devils, we set the value of  $\sigma_{\text{th}}$  individually for each combination, as shown in Table A1. The selected  $\sigma_{\text{th}}$  values are used consistently to evaluate hallucination performance throughout all experiments in the main paper. As described in Section 4.1, PGD-based adversarial attacks are performed with  $k = 3$  and 200 iterations. For uncertainty estimation, masks  $M$  are extracted from layers  $\mathcal{S} = \{1, \dots, 10\}$  of the vision encoder. The masking operation is applied within the self-attention mechanism of the vision encoder, targeting layers 13–17 for LLaVA-1.5 and Shikra, and layers 9–16 for MiniGPT-4. All experiments in the main paper were conducted on an NVIDIA A100 GPU with 80GB of memory.

Table A1: **Values of  $\sigma_{\text{th}}$  for each model and method combination.** We determine  $\sigma_{\text{th}}$  individually for each combination and use the selected value consistently across all evaluations to ensure fair and robust comparisons.

| Model         | Greedy | OPERA | VCD | PAI | Devils |
|---------------|--------|-------|-----|-----|--------|
| LLaVA-1.5-7B  | 1.1    | 1.1   | 1.0 | 1.8 | 1.9    |
| LLaVA-1.5-13B | 1.2    | 1.2   | 1.1 | 1.6 | 1.6    |
| Shikra-7B     | 1.0    | 1.0   | 1.0 | 1.5 | 1.9    |
| MiniGPT-4     | 0.0    | 2.0   | 0.0 | 1.1 | -0.1   |

## D Experimental Details

### D.1 Benchmarks

**CHAIR.** To evaluate the robustness of image captioning models against object hallucination, we adopt the CHAIR [47] metric (Caption Hallucination Assessment with Image Relevance). This benchmark quantifies hallucination by comparing generated captions with ground truth object annotations and sentence descriptions in the MSCOCO dataset. Two variants,  $\text{CHAIR}_i$  and  $\text{CHAIR}_s$ , measure hallucination at the object and sentence levels, respectively, as shown in Eq. 5.

This metric enables a systematic comparison of hallucination severity across models and offers insights into the alignment between visual input and generated language beyond standard evaluation metrics. We use the prompt “Please describe this image in detail.”.

**POPE.** To obtain a more reliable and instruction-agnostic assessment of object hallucination in large vision-language models (LVLMs), we adopt the POPE (Polling-based Object Probing Evaluation) framework [31]. Unlike traditional caption-based metrics that are sensitive to prompt phrasing and rely on manual parsing, POPE probes a model’s visual grounding through binary yes/no questions about object presence. This enables stable and scalable evaluation across both annotated and unannotated datasets. POPE constructs evaluation sets using three sampling strategies: Random, Popular, and Adversarial. Each strategy targets a different source of hallucination, allowing us to test whether models tend to hallucinate arbitrary objects, frequently occurring objects, or objects that often co-occur with those actually present in the image. We use the prompt “Is there a/an [object] in the image?”.

**AMBER.** To evaluate object hallucination comprehensively in large vision-language models (LVLMs), we adopt the AMBER benchmark [56]. AMBER assesses hallucinations across both generative and discriminative tasks, focusing on three primary types: existence, attribute, and relation. In the generative setting, it employs metrics such as CHAIR, Hal, and Cog to measure hallucination



frequency, object coverage, and cognitive tendencies. For discriminative tasks, standard binary classification metrics are used, and the AMBER Score integrates CHAIR from the generative setting with the F1 score from the discriminative setting. Notably, we focus exclusively on ‘existence’ subset to assess object hallucination, which involves generating descriptions of objects that are not present in the input image. We use the prompt “Describe this image.” for generative task and “Is there a [object] in this image?” for discriminative task.

## D.2 Base models

**LLaVA-1.5.** In our experiments, we employed LLaVA-1.5 [35], a versatile multimodal model developed for visual instruction tuning. LLaVA-1.5 builds upon the original LLaVA [36] architecture by integrating a two-layer MLP as a vision-language connector, leveraging the CLIP-ViT-L-336px [46] vision encoder, and incorporating academic task-oriented VQA data with response formatting prompts. These modifications significantly enhance the model’s capability for both visual reasoning and instruction following, while retaining strong data efficiency. LLaVA-1.5 achieves competitive performance across a broad set of multimodal benchmarks using only publicly available data and modest computational resources. To investigate the robustness of our method across different model scales, we conducted experiments using both the 7B and 13B versions of LLaVA-1.5. This enabled us to evaluate whether our approach maintains performance consistency under varying model capacities. For the experiments, we utilized the official implementation <sup>1</sup> along with the provided code and model weights.

**Shikra.** In our experiments, we adopt the Shikra-7B [9] model, a LVLm specifically designed for referential dialogue. Shikra-7B integrates a CLIP-ViT-L/14 [46] vision encoder with a Vicuna-7B language model via a simple alignment layer, allowing end-to-end processing without the need for additional vocabularies, position encoders, detection modules, or external plug-ins. A key feature of Shikra is its ability to represent spatial information directly in natural language using numerical coordinates, allowing it to handle both inputs and outputs involving region references seamlessly. This architecture supports a broad range of vision-language tasks, including Visual Question Answering (VQA), image captioning, referring expression comprehension (REC), and PointQA, all within a unified framework and without task-specific fine-tuning. Its strong performance across both conventional and location-sensitive tasks makes it a compelling choice for measuring object hallucination. For the experiments, we utilized the official implementation <sup>2</sup> along with the provided code and model weights.

**MiniGPT-4.** In our experiments, we employed MiniGPT-4 [69] as a vision-language model to evaluate effectiveness of our method. MiniGPT-4 combines a frozen vision encoder from BLIP-2 [30] (EVA-CLIP-ViT-G/14 [51] with Q-Former) and a large frozen language model, Vicuna, using a single trainable linear projection layer to align visual features with the input space of the language model. The model is pre-trained on approximately 5 million image-text pairs to establish initial multimodal capabilities. To address issues such as repetitive or fragmented outputs observed after pretraining, a second stage fine-tuning is applied using a curated set of 3,500 detailed image-description pairs, formatted with a conversational prompt template. This two-stage training strategy improves the fluency and relevance of the model’s responses, enabling it to handle a variety of vision-language tasks more effectively. When applying our methodology to MiniGPT-4, we conducted the adversarial attack on the features prior to their input into the Q-Former. For the experiments, we utilized the official implementation <sup>3</sup> along with the provided code and model weights.

## D.3 Baselines

**Greedy.** Greedy decoding is one of the most basic decoding strategies for generative language models, where the token with the highest prediction probability is selected at each step. This approach is fast and straightforward to implement. Among various decoding strategies for LVLms, we adopt the naïve and fundamental greedy decoding method as one of our baselines to evaluate the object hallucination mitigation performance of our method.

<sup>1</sup><https://github.com/haotian-liu/LLaVA>

<sup>2</sup><https://github.com/shikras/shikra>

<sup>3</sup><https://github.com/Vision-CAIR/MiniGPT-4>

Table A2: **Runtime comparison between MC dropout and our method using PGD-based adversarial attack.** When comparing the mean runtime, our method is  $\times 5.1$  faster. The symbol  $\pm$  denotes the  $1\sigma$  interval.

| Method   | MC dropout          | Adversarial attack (Ours) |
|----------|---------------------|---------------------------|
| Time (s) | 12.4 ( $\pm 0.12$ ) | 2.43 ( $\pm 0.08$ )       |

**OPERA.** The authors of OPERA [20] identify that object hallucination in LVLMs is closely linked to specific knowledge aggregation patterns within the model’s self-attention matrix. It defines tokens that induce such attention patterns as summary tokens and mitigates hallucination by detecting excessive attention toward these tokens and preventing their influence on next-token prediction. Specifically, OPERA extracts a local window from the self-attention map, quantifies the degree of aggregation via column-wise multiplication, and applies a logit penalty during beam search to suppress over-confident candidates. While effective, OPERA relies on beam search, which introduces significant additional computational cost. For comparison and integration with our method, we used the official implementation <sup>4</sup> provided by the authors.

**Visual Contrastive Decoding.** The authors of Visual Contrastive Decoding (VCD) [45] attribute object hallucination to statistical biases, such as object cooccurrence frequencies in training data, and language priors inherent to large language models. By injecting Gaussian noise into the input image, the LVLM’s reliance on visual information is reduced, causing it to lean more heavily on these language priors. To counteract this, VCD introduces both the original image  $v$  and a distorted version  $v'$  as input, computes their respective output probability distributions, and then extrapolates a contrastive probability distribution that suppresses language-driven biases. For comparison and integration with our method, we use the official implementation <sup>5</sup>. When applying our method to VCD, we performed uncertain token suppression only on the original image  $v$ .

**PAI.** The authors of Paying more Attention to Image (PAI) [37] argue that object hallucination arises when visual information is ignored and propose a training-free method to enhance the influence of images during inference. Specifically, they manipulate the self-attention matrix to amplify attention toward visual tokens and selectively strengthen particular attention heads to guide the model toward more trustworthy directions. To avoid excessive attention toward the beginning-of-sentence (BOS) token, they introduce a layer prior that excludes shallow layers from modulation. Additionally, they compare outputs with and without the input image to attenuate language model biases. Since PAI does not modify the vision encoder, our method can be additionally applied. For comparison, we utilized the official implementation <sup>6</sup>.

**Devils in the middle layers.** In Devils in the Middle Layers (Devils) [24], the authors find that in large vision-language models (LVLMs), visual information is strongly processed in the middle layers of the language model. They observe that inactive attention can induce hallucinations, and that during such instances, attention heads tend to focus inconsistently on unrelated objects. To address this, the authors propose integrating information across attention heads during inference to encourage focus on more consistent visual regions. They achieve this by reweighting the attention scores to emphasize coherent areas. Since this is an intervention on the LLM component, their methodology is applicable in our setting as well. To implement it, we adopted their official codebase <sup>7</sup>.

## E Additional Analysis

### E.1 Monte Carlo vs. Adversarial attack

In the main paper, we verify the similarity between the uncertainty map  $U$  obtained via adversarial attacks and the one derived from the Monte Carlo (MC) dropout using pre-trained vision encoder. To

<sup>4</sup><https://github.com/shikiw/OPERA>

<sup>5</sup><https://github.com/DAMO-NLP-SG/VCD>

<sup>6</sup><https://github.com/LALBJ/PAI>

<sup>7</sup>[https://github.com/ZhangqiJiang07/middle\\_layers\\_indicating\\_hallucinations](https://github.com/ZhangqiJiang07/middle_layers_indicating_hallucinations)

Table A3: **Object hallucination benchmark results under varying attack strengths** ( $\|\epsilon\|_\infty$ ). To investigate the effect of adversarial perturbations on the image encoder, we applied PGD attacks of different magnitudes for 200 iterations to LLaVA-1.5-7B and evaluated performance using the CHAIR benchmark. Adversarial attacks on the image encoder increase the likelihood of hallucinated outputs, with the severity of hallucination correlating positively with the attack strength.

| $\ \epsilon\ _\infty$ | CHAIR <sub>s</sub> ↓ | CHAIR <sub>i</sub> ↓ | Recall ↑ | Precision ↑ | F1 ↑ |
|-----------------------|----------------------|----------------------|----------|-------------|------|
| 0                     | 47.4                 | 12.2                 | 78.9     | 76.9        | 77.9 |
| 1                     | 53.0                 | 16.2                 | 76.9     | 72.9        | 74.8 |
| 3                     | 64.0                 | 25.5                 | 63.0     | 62.4        | 62.7 |
| 5                     | 65.6                 | 25.9                 | 55.9     | 60.1        | 57.9 |
| 7                     | 61.6                 | 26.6                 | 50.5     | 59.6        | 54.7 |

further confirm this similarity, we provide an additional qualitative comparison in Fig. A1. Although our method tends to slightly overestimate the uncertainty, it consistently identifies high-uncertainty regions that closely align with those highlighted by MC dropout. To assess the computational efficiency of our approach, we compare the runtime of uncertainty estimation using Monte Carlo dropout and our adversarial-based method. Specifically, we apply both techniques to the vision encoder from LLaVA-1.5-7B. The adversarial attack is performed 100 times with  $k = 3$  top perturbations, while the Monte Carlo dropout requires 1,000 forward passes, both executed on a single NVIDIA RTX 4090 GPU. The results, presented in Table A2, demonstrate that our method enables significantly more efficient extraction of uncertainty masks, highlighting its practical advantage in identifying visually uncertain tokens.

## E.2 Effect of adversarial attacks on LVLM outputs

We conducted PGD-based adversarial attacks on the vision encoder to identify the uncertain visual tokens. To evaluate whether such attacks effectively influence the output of LVLMs, we applied adversarial perturbations with varying magnitudes of  $\epsilon$  and performed both quantitative and qualitative analyses.

As shown in Fig. A2, the responses generated from the attacked images often exhibited hallucinations or failed to produce correct answers. As demonstrated in Table A3, we also observe that higher attack intensities lead to increased severity of hallucinations. These experimental results highlight that the visual features extracted by the vision encoder play a crucial role in LVLMs’ performance of downstream task, emphasizing that enhancing visual perception is critical for reducing hallucination and improving overall reliability.

## E.3 Consistency and robustness of uncertainty masks from adversarial attacks

We identify uncertain visual tokens by applying PGD-based adversarial attacks to the features of the vision encoder. In our implementation, the attack is initialized from the original image without added noise. To evaluate the consistency and robustness of the resulting uncertainty masks  $M$ , we also perform attacks with different initial noise seeds, generating diverse adversarial perturbations. From each perturbed image, we extract a mask and compute the mean Intersection over Union (mIoU) between the masks  $M$  generated from different seeds.

As shown in Table A4, the uncertainty masks  $M$  remain highly consistent across different initializations. Qualitative examples in Fig. A3 further demonstrate that the uncertainty maps  $U$  and masks  $M$  maintain stable and coherent structures. These results confirm the reliability of our method in consistently identifying uncertain tokens under varying adversarial conditions.

## F Additional Ablation Studies

**Masking Threshold Hyperparameter  $\sigma_{th}$ .** To construct the binary uncertainty mask  $M$ , we introduce a threshold hyperparameter  $\sigma_{th}$ . Its optimal value depends on the characteristics of each model and method combination, and is determined through grid search. Table A5 presents an ablation study conducted on the LLaVA-1.5-7B model using six different threshold values. Considering the

Table A4: **Mask consistency measured by mean Intersection over Union (mIoU).** We applied adversarial attacks to the LLaVA-1.5-7B image encoder on 500 images across five different seeds and measured the mIoU to verify mask consistency. The results indicate that the masks obtained through adversarial attack are robust and consistent. The threshold  $\sigma_{th}$  was set to 1.1.

| Seed pair | (0, 1)                | (0, 2)                | (0, 3)                | (0, 4)                | (1, 2)                |
|-----------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| mIoU      | 0.899 ( $\pm 0.034$ ) | 0.898 ( $\pm 0.035$ ) | 0.898 ( $\pm 0.036$ ) | 0.899 ( $\pm 0.036$ ) | 0.899 ( $\pm 0.035$ ) |
| Seed pair | (1, 3)                | (1, 4)                | (2, 3)                | (2, 4)                | (3, 4)                |
| mIoU      | 0.898 ( $\pm 0.036$ ) | 0.898 ( $\pm 0.035$ ) | 0.897 ( $\pm 0.036$ ) | 0.897 ( $\pm 0.036$ ) | 0.897 ( $\pm 0.036$ ) |

Table A5: **Ablation study of the thresholding parameter  $\sigma_{th}$  for generating the uncertainty mask  $M$ .** We use LLaVA-1.5-7B with greedy decoding and evaluate hallucination performance while varying the threshold  $\sigma_{th}$ .

| $\sigma_{th}$    | Greedy | 0.8  | 0.9  | 1.0  | 1.1  | 1.2  | 1.3  |
|------------------|--------|------|------|------|------|------|------|
| $C_s \downarrow$ | 47.4   | 27.0 | 27.0 | 30.0 | 29.2 | 33.6 | 36.4 |
| $C_i \downarrow$ | 12.2   | 8.4  | 8.2  | 9.0  | 9.3  | 9.7  | 10.3 |
| F1 $\uparrow$    | 77.9   | 76.7 | 77.7 | 77.6 | 78.2 | 78.0 | 78.5 |

trade-offs among  $C_s$ ,  $C_i$ , and F1 score, we select  $\sigma_{th} = 1.1$  as it yields the best overall performance. Based on this analysis, we apply the optimal  $\sigma_{th}$  for each configuration in our experiments.

## G Additional Quantitative and Qualitative Results

### G.1 Computational Cost

Our method identifies uncertain tokens via PGD-based adversarial attacks implemented through backpropagation, which naturally introduces additional computational overhead compared to standard greedy decoding. To quantify this cost, we measure the extra inference time and compare it with existing hallucination mitigation methods. As shown in Table A6, while our method does incur some additional overhead, it offers comparable or even lower inference time than several baselines, achieving a favorable balance between performance and efficiency.

### G.2 Additional quantitative results

**Applicability of our method to larger model.** We assess the scalability and generalizability of our method using the larger LLaVA-1.5-13B model. As shown in Table A7, our method delivers substantial improvements over the greedy decoding baseline, reducing  $C_s$  by 15.2 and  $C_i$  by 2.9. It also integrates effectively with a variety of existing approaches, achieving the best performance when combined with Devils ( $C_s = 20.4$ ,  $C_i = 6.0$ ). These results demonstrate that our method generalizes well across model scales and enhances a wide range of existing hallucination mitigation strategies.

Table A6: **Additional inference time introduced by each method compared to standard greedy decoding.** We performed text generation with request of image description with max 32 tokens. All experiments were conducted using LLaVA-1.5-7B on an NVIDIA A100 GPU. We report the mean and standard deviation over 30 samples. Although our method introduces some overhead due to backpropagation from PGD attacks, it remains comparable to or even faster than existing approaches.

| Method | Additional inference time (s) |
|--------|-------------------------------|
| OPERA  | 9.518 $\pm$ 0.011             |
| VCD    | 1.646 $\pm$ 0.001             |
| PAI    | 1.567 $\pm$ 0.021             |
| Devils | 0.014 $\pm$ 0.001             |
| Ours   | 2.469 $\pm$ 0.004             |

Table A7: **Quantitative results on CHAIR benchmark for LLaVA-1.5-13B.** We report object hallucination ( $C_s$ ,  $C_i$ ) for various mitigation methods and their combination with our method. The maximum token length is set to 512.  $\Delta\%$  denotes the relative improvement in performance.

| Method           | Greedy      |       |                    | OPERA |       |                    | VCD   |       |                    | PAI   |             |                   | Devils |             |                    |
|------------------|-------------|-------|--------------------|-------|-------|--------------------|-------|-------|--------------------|-------|-------------|-------------------|--------|-------------|--------------------|
|                  | Orig.       | +Ours | $\Delta\%$         | Orig. | +Ours | $\Delta\%$         | Orig. | +Ours | $\Delta\%$         | Orig. | +Ours       | $\Delta\%$        | Orig.  | +Ours       | $\Delta\%$         |
| $C_s \downarrow$ | 45.4        | 30.2  | $\uparrow 33.4\%$  | 40.2  | 30.4  | $\uparrow 24.4\%$  | 49.0  | 35.4  | $\uparrow 27.8\%$  | 38.6  | 32.4        | $\uparrow 16.1\%$ | 28.2   | <b>20.4</b> | $\uparrow 26.2\%$  |
| $C_i \downarrow$ | 11.2        | 8.3   | $\uparrow 25.9\%$  | 10.9  | 8.9   | $\uparrow 18.3\%$  | 13.4  | 10.3  | $\uparrow 23.1\%$  | 9.9   | 8.4         | $\uparrow 15.2\%$ | 8.7    | <b>6.0</b>  | $\uparrow 31.0\%$  |
| F1 $\uparrow$    | <b>79.1</b> | 78.9  | $\downarrow 0.2\%$ | 78.0  | 76.9  | $\downarrow 1.4\%$ | 77.3  | 76.3  | $\downarrow 1.3\%$ | 78.7  | <b>79.1</b> | $\uparrow 0.3\%$  | 78.4   | 78.0        | $\downarrow 0.5\%$ |

Table A8: **Quantitative results of our method on state-of-the-art LVLMS.** We apply our approach to two SOTA models, DeepSeek-VL and Qwen2.5-VL, and compare performance against greedy decoding. For DeepSeek-VL we set  $\sigma_{th} = 1.0$ , while for Qwen2.5-VL we use  $\sigma_{th} = 0.0$ . These results demonstrate that our method is applicable to a wide range of LVLMS, including the most recent architectures.

| Method               | CHAIR            |                  |               | POPE        |             |             |
|----------------------|------------------|------------------|---------------|-------------|-------------|-------------|
|                      | $C_s \downarrow$ | $C_i \downarrow$ | F1 $\uparrow$ | Rand.       | Pop.        | Adv.        |
| DeepSeek-VL (Greedy) | 25.8             | 6.6              | <b>72.7</b>   | 88.7        | 88.0        | 84.9        |
| +Ours                | <b>22.4</b>      | <b>5.5</b>       | 72.6          | <b>88.8</b> | <b>88.0</b> | <b>85.1</b> |
| Qwen2.5-VL (Greedy)  | 29.6             | 7.8              | 76.0          | 84.2        | 83.7        | 83.3        |
| +Ours                | <b>28.6</b>      | <b>7.0</b>       | <b>76.8</b>   | <b>84.3</b> | <b>83.8</b> | <b>83.4</b> |

Table A9: **Additional quantitative results for an alternative adversarial attack on a Q-Former-based LVLMS architecture.** MiniGPT-4 uses a Q-Former to effectively compress image tokens, which confers robustness to image-only perturbations. By jointly perturbing the Q-Former’s learnable query vectors together with the image, we enable a stronger attack and observe additional gains in attack effectiveness.

| Method                | $C_s \downarrow$ | $C_i \downarrow$ | F1 $\uparrow$ |
|-----------------------|------------------|------------------|---------------|
| Greedy (MiniGPT-4)    | 31.0             | 11.4             | 67.3          |
| +Ours (Image only)    | 29.0             | 10.6             | 67.5          |
| +Ours (Image + Query) | <b>27.0</b>      | <b>9.3</b>       | <b>68.1</b>   |

**Applicability of our method to the state-of-the-art models.** In the main paper, we conducted extensive experiments on LLaVA-1.5, Shikra, and MiniGPT, which are commonly used as target models in object hallucination mitigation studies and therefore served as our primary evaluation benchmarks. To further validate the applicability of our approach, we additionally evaluated state-of-the-art models such as DeepSeek-VL [38] and Qwen2.5-VL [4]. These models not only demonstrate strong performance, but also involve joint fine-tuning of the vision encoder during vision-language alignment training, making them suitable indicators of the scalability of our method. The results presented in table A8 confirm that our approach effectively reduces object hallucination even in these latest models.

**Alternative attack methods on Q-Former design architecture.** We observed that adversarial attacks applied solely to the image have limited effectiveness in Q-Former based architectures (*e.g.*, MiniGPT-4). This appears to stem from the robustness introduced by the architectural design that relies on learnable queries. To validate this hypothesis, we additionally optimized the input queries during adversarial attacks to examine whether our approach provides further advantages. Unlike images, the query vectors are continuous, and thus we imposed a noise constraint on the query vector  $q$  such that the perturbation scale matches that applied to the image.

$$\|\epsilon_q\|_\infty = \frac{\|\epsilon\|_\infty}{255} \cdot \frac{(\max(q) - \min(q))}{2}, \quad (\text{A13})$$

where  $\epsilon_q$  is the adversarial noise injected to query vectors  $q$ ,  $\epsilon$  is the noise added to the victim image. The results are presented in table A9, which report the outcomes of adversarial attacks jointly applied

Table A10: **Average length of generated text with standard deviation.** We report the average length of generated texts across different models and hallucination mitigation methods, with and without our approach. Values are presented as mean  $\pm$  standard deviation. Our method slightly reduces output length, which has been linked to lower hallucination rates in LVLMs.

| Model     | Greedy        |               | OPERA         |               | VCD           |               | PAI           |               | Devils        |               |
|-----------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
|           | Orig.         | +Ours         | Orig.         | +Ours         | Orig.         | +Ours         | Orig.         | +Ours         | Orig.         | +Ours         |
| LLaVA-7B  | 491 $\pm$ 104 | 426 $\pm$ 105 | 473 $\pm$ 107 | 406 $\pm$ 118 | 517 $\pm$ 114 | 420 $\pm$ 121 | 514 $\pm$ 118 | 487 $\pm$ 120 | 504 $\pm$ 206 | 448 $\pm$ 173 |
| LLaVA-13B | 495 $\pm$ 101 | 440 $\pm$ 114 | 452 $\pm$ 136 | 402 $\pm$ 142 | 515 $\pm$ 108 | 436 $\pm$ 126 | 510 $\pm$ 122 | 468 $\pm$ 115 | 406 $\pm$ 141 | 381 $\pm$ 124 |
| Shikra-7B | 514 $\pm$ 110 | 475 $\pm$ 108 | 370 $\pm$ 120 | 354 $\pm$ 109 | 524 $\pm$ 113 | 487 $\pm$ 113 | 493 $\pm$ 195 | 427 $\pm$ 213 | 383 $\pm$ 202 | 368 $\pm$ 265 |
| MiniGPT-4 | 408 $\pm$ 206 | 418 $\pm$ 202 | 301 $\pm$ 135 | 304 $\pm$ 110 | 404 $\pm$ 167 | 404 $\pm$ 172 | 284 $\pm$ 126 | 282 $\pm$ 130 | 415 $\pm$ 444 | 391 $\pm$ 389 |

Table A11: **Effectiveness of our method applied to different decoding baselines.** We evaluate our method on LLaVA-1.5-7B using various decoding strategies, including greedy decoding, beam search, DoLa and VAR. We set the  $N_{beam} = 5$ . Across all settings, our method consistently reduces hallucination metrics ( $C_s$ ,  $C_i$ ) while maintaining or improving F1 score.

| Method      | $C_s \downarrow$ | $C_i \downarrow$ | F1 $\uparrow$ |
|-------------|------------------|------------------|---------------|
| Greedy      | 47.4             | 12.2             | 77.9          |
| +Ours       | 29.2             | 9.3              | 78.2          |
| Beam search | 47.2             | 12.7             | 77.8          |
| +Ours       | 28.2             | 8.6              | 78.5          |
| DoLa        | 46.0             | 12.2             | 78.5          |
| +Ours       | 30.4             | 9.5              | 78.2          |
| VAR         | 46.8             | 12.5             | 77.9          |
| +Ours       | 29.4             | 9.1              | 78.1          |

Table A12: **Comparison of uncertainty estimation methods for generating mask  $M$ .** We evaluate the effectiveness of our adversarial attack-based uncertainty estimation method against MC dropout on LLaVA-1.5-7B using the CHAIR dataset.

| Method                | $C_s \downarrow$ | $C_i \downarrow$ | F1 $\uparrow$ |
|-----------------------|------------------|------------------|---------------|
| Greedy                | 47.4             | 12.2             | 77.9          |
| +Ours (w/Adv. attack) | 29.2             | 9.3              | 78.2          |
| +Ours (w/MC dropout)  | 32.6             | 10.5             | 77.8          |

to both the image and the Q-Former queries. The evaluation on the CHAIR benchmark demonstrates that our method can achieve further performance improvements when combined with additional architectural considerations. However, for methodological consistency, the main paper focuses only on adversarial perturbations applied to the image.

**Length of generated text.** [64] highlights that overly long outputs from LVLMs often lead to object hallucinations, as the generated content exceeds the model’s visual perception. As shown in Table A10, our method consistently and slightly reduces the length of image descriptions across various models and hallucination mitigation methods. However, in the case of MiniGPT-4, due to its Q-Former architecture, masking uncertain visual tokens within the vision encoder is less effective. As a result, the generated text length may occasionally remain unchanged or even slightly increase.

**Application of our method to other baselines.** To validate the generalizability of our method for mitigating object hallucination in LLaVA-1.5-7B, we apply it to alternative decoding strategies, including beam search decoding [52], DoLa [12] and VAR [25], using the CHAIR dataset. As shown in Table A11, our method consistently reduces hallucination rates while maintaining or even improving the F1 score.

**Comparison of uncertainty estimation of visual token: Our Method vs. MC Dropout.** Epistemic uncertainty of visual tokens introduced by a pre-trained vision encoder can be estimated using MC Dropout. However, this approach often requires intensive computation due to thousands of forward passes. As an efficient alternative, we propose a method that estimates uncertainty of visual tokens using PGD-based adversarial attacks.

We perform experiments on LLaVA-1.5-7B using the CHAIR dataset and compare the uncertainty masks  $M$  for visual tokens, generated using Eq.3, between our method and MC Dropout. As shown

in Table A12, our approach achieves comparable or better performance while being more computationally efficient. These results highlight that our PGD-based uncertainty estimation effectively captures the epistemic uncertainty of the pre-trained vision encoder and reliably identifies uncertain visual tokens.

Regarding the lower performance of MC dropout compared to our method, we conjecture that although MC dropout is widely used for uncertainty quantification, it remains only one estimation technique. In contrast, our approach provides a more conservative estimate of uncertainty through an upper bound, which we believe accounts for its superior performance.

### G.3 Additional qualitative results

**Qualitative examples of binary uncertainty masks  $M$ .** Fig. A4 presents additional examples of binary uncertainty masks  $M$  generated for various input images under PGD-based adversarial attacks applied to the vision encoder of LLaVA-1.5-7B.

**Qualitative examples of our method on various LVLMS with different mitigation methods.** We present additional qualitative examples of our method applied to different combinations of LVLMS (LLaVA-1.5-7B and Shikra-7B) and hallucination mitigation techniques, including greedy decoding, OPERA, VCD, PAI, and Devils. Our method integrates well with these approaches and effectively reduces object hallucinations by preventing the generation of non-existent objects. Fig. A5–A24 illustrate qualitative examples on the CHAIR and POPE datasets using LLaVA-1.5-7B and Shikra-7B across various hallucination mitigation methods.

**Qualitative examples of failure cases.** Fig. A25 presents qualitative examples of failure cases from our proposed method. Although our method consistently mitigates hallucinated responses, it occasionally fails to prevent all hallucinations.

## H Discussion

We statistically demonstrate that epistemic uncertainty within the vision encoder contributes to object hallucination and address this issue through self-attention masking at intermediate layers. To understand how LVLMS change their integration of visual information after applying our method, we measured the entropy of the LLM’s attention distribution over image tokens across all layers and heads. Entropy serves as an indicator of whether the model attends broadly or narrowly, with higher entropy reflecting the use of a wider range of visual evidence rather than reliance on a small subset of tokens. Using 500 images, we found that the average entropy of LLaVA increased from 1.5746 in the original model to 1.9717 with our method. This increase suggests that our approach encourages broader and more balanced attention over reliable visual tokens, enabling the model to integrate visual information more effectively while reducing over-reliance on uncertain inputs, consistent with findings from prior work [37].

## I Broader Impacts

We proposed a method to improve the reliability of Large Vision-Language Models (LVLMS) by identifying and masking uncertain visual tokens in the vision encoder, a key source of object hallucination. In contrast to existing approaches that intervene at the language model level, our method operates solely on the vision encoder and demonstrates effectiveness across a variety of models and settings.

Our method offers significant societal benefits by improving safety and reliability in critical applications such as medical imaging, assistive technologies, and autonomous systems. However, it may also inadvertently suppress valid but ambiguous visual information, which could disproportionately affect underrepresented groups and reinforce existing dataset biases, raising important concerns about potential negative societal impacts.



## J Limitations

Despite its advantages, our method has several limitations. First, while it effectively reduces hallucinations, the proposed masking strategy within self-attention process of vision encoder may result in the loss of visual information, leading to slight performance degradation in certain perception-oriented tasks. Second, the reliance on PGD-based adversarial attacks to estimate uncertainty increases inference time. Third, the masking strategy lacks a formal theoretical foundation, instead relying on a trace-based approximation of uncertainty as a surrogate for the determinant of the covariance matrix. Finally, the method is less effective for models like MiniGPT-4, which utilize a Q-Former to abstract visual information before integrating it with language. In these cases, masking at the vision encoder has limited downstream impact. Addressing these issues is a promising direction for future research.

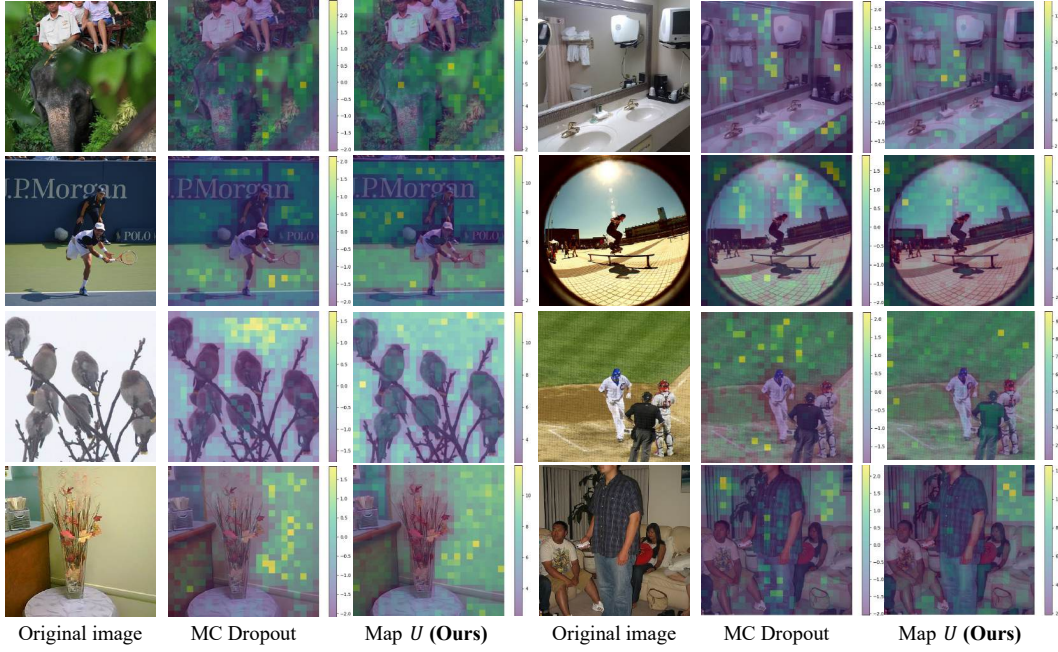


Figure A1: **Additional visual comparison of estimated uncertainty from MC dropout and our method.** Our uncertainty map  $U$  identifies uncertain regions similar to the uncertainty map obtained via MC dropout. MC dropout was applied to the self-attention and MLP residuals of the LLaVA-1.5 vision encoder with a dropout rate of  $p = 0.5$ . Layers  $i = 5$  to  $j = 8$  were used, and the variance of each token was estimated over 1,000 forward passes. For the adversarial attack, we applied 100 iterations of PGD with  $k = 3$ . The MC-based uncertainty values were log-scaled for visualization clarity.



Figure A2: **Qualitative comparison of LVLM outputs when using the original image versus the adversarially attacked image as input.** When conditioned on the attacked image (right column), the model generates descriptions involving non-existent objects or scenes, indicating a more severe object hallucination effect compared to the original image (left column). This experiment was conducted on the LLaVA-1.5-7B model using the CHAIR dataset, with  $k = 3$  and 200 PGD iterations.

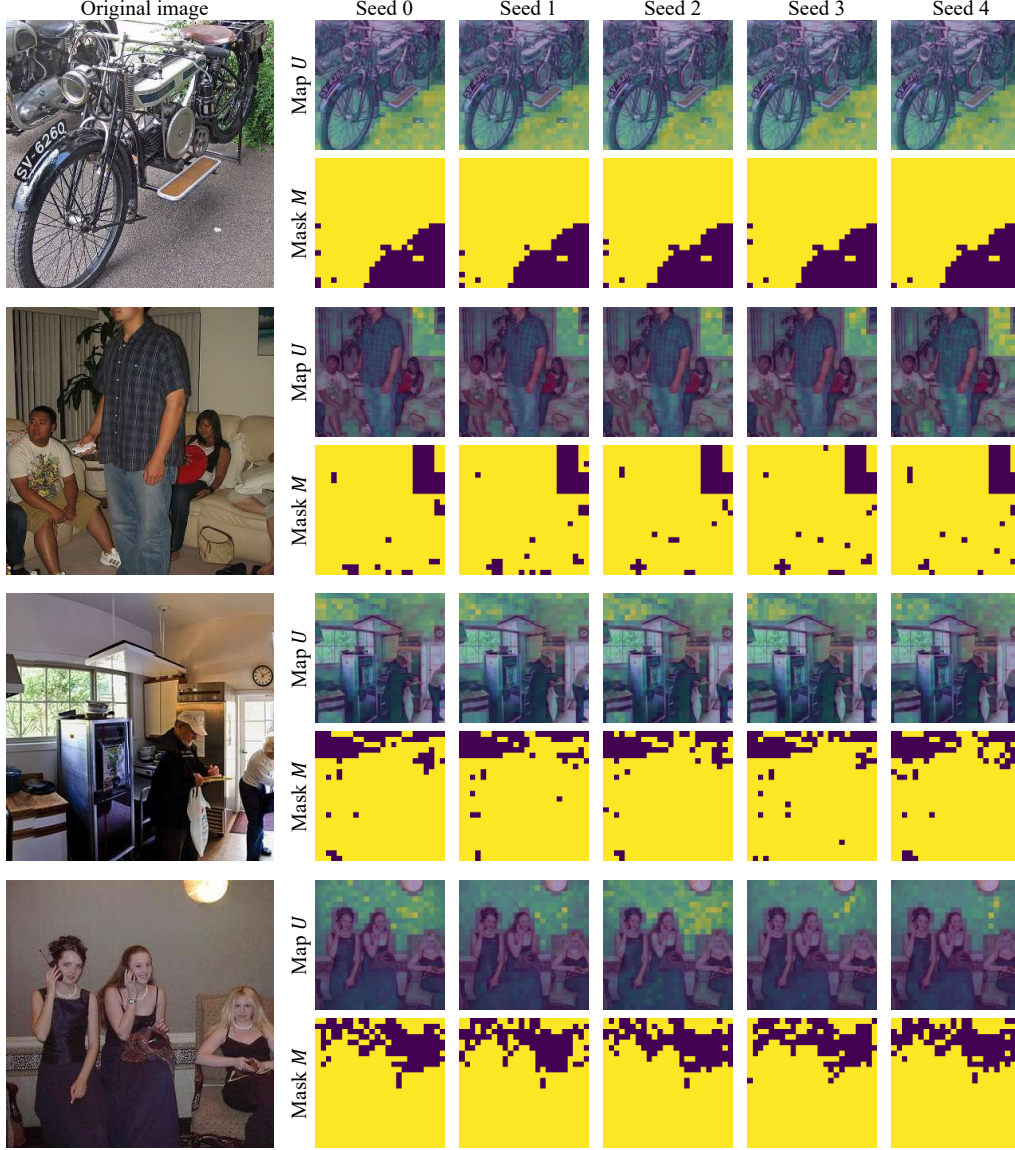


Figure A3: **Uncertainty maps  $U$  and masks  $M$  generated from different initial noises using PGD-based adversarial attacks.** We qualitatively demonstrate the consistency of the uncertainty maps  $U$  and corresponding masks  $M$  ( $\sigma_{th} = 1.1$ ), generated using PGD-based adversarial attacks with five different random seeds for initializing noise. Despite variations in the initial noise, the resulting uncertainty maps  $U$  and masks  $M$  remain highly similar, highlighting the robustness and stability of the attack-based uncertainty estimation in LLaVA-1.5-7B vision encoder.

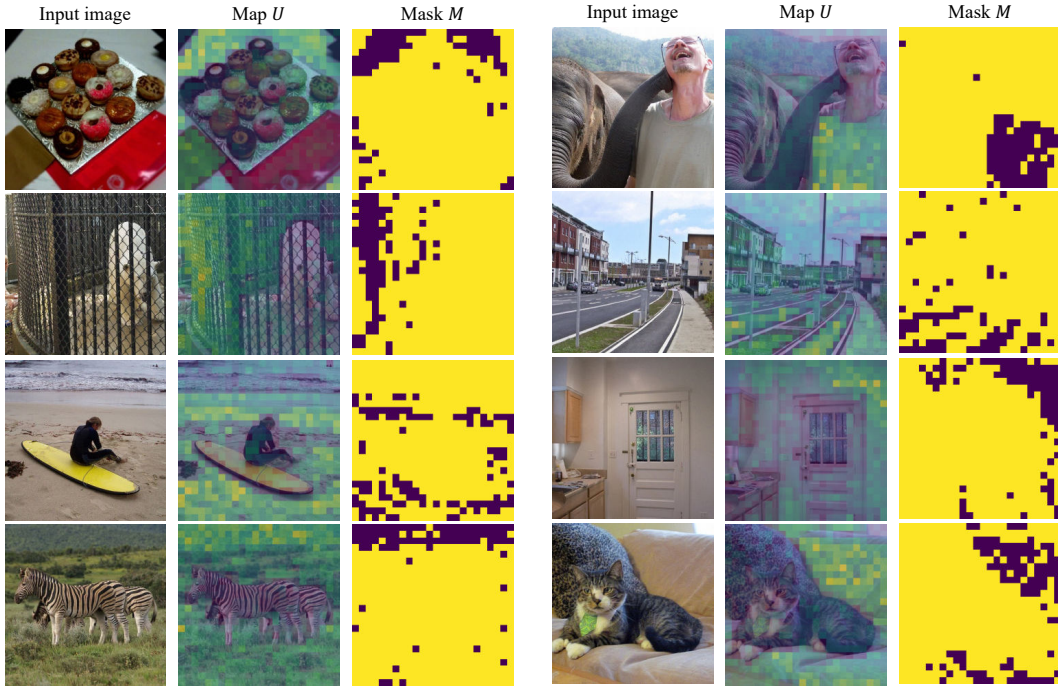


Figure A4: **Additional qualitative examples of uncertainty maps  $U$  and masks  $M$  using PGD-based adversarial attacks.** We visualize the uncertainty maps  $U$  and the corresponding masks  $M$  (with  $\sigma_{\text{th}} = 1.1$ ) generated from adversarially perturbed inputs across various samples. All examples were obtained using the LLaVA-1.5-7B vision encoder.



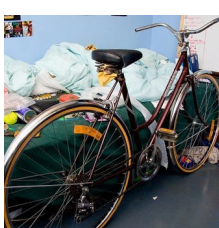
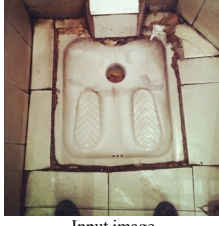
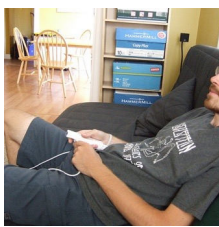
|                                                                                                        |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |                                                                                                      |
|--------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------|
| <br>USER              | Please describe the image in detail.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | <br>Input image   |
| <br>LLaVA             | <p>The image features a bathroom with a white sink and a mirror above it. The sink is situated next to a shower, which is covered in newspaper pages. The walls of the bathroom are adorned with a unique design, featuring a combination of newspaper and tile. There are several towels in the bathroom, with one placed on the sink and another on the floor. Additionally, there are <b>two bottles</b>, one near the sink and the other on the floor. A <b>cup</b> can also be seen on the sink, and a <b>book</b> is placed on the floor.</p>                                                                                |                                                                                                      |
| <br>LLaVA<br>+ Ours   | <p>The image features a bathroom with a sink and a mirror. The sink is located on the left side of the bathroom, while the mirror is positioned above it. The bathroom also has a shower, which is situated on the right side of the room. The walls of the bathroom are adorned with a unique and interesting design, featuring a collage of newspaper clippings. This artistic choice adds a touch of personality and character to the space.</p>                                                                                                                                                                                |                                                                                                      |
| <br>USER              | Please describe the image in detail.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | <br>Input image   |
| <br>LLaVA             | <p>The image features a large white and green bus driving down a city street. The bus is positioned in the middle of the street, and it appears to be a public transit bus. There are several people walking along the sidewalk, with some of them carrying <b>handbags</b>. In addition to the bus, there are two bicycles parked on the sidewalk, one closer to the left side of the image and the other further to the right. A <b>traffic light</b> can be seen on the left side of the street, and a <b>fire hydrant</b> is located near the center of the scene.</p>                                                         |                                                                                                      |
| <br>LLaVA<br>+ Ours   | <p>The image shows a large white bus driving down a street. The bus is a public transit bus, and it is driving down the road. The bus is in the middle of the street, and it is surrounded by buildings. The bus is also driving past a bus stop, which is located on the left side of the street. The bus is the main focus of the image, and it is driving down the road, providing transportation services to the people in the area.</p>                                                                                                                                                                                       |                                                                                                      |
| <br>USER              | Please describe the image in detail.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | <br>Input image  |
| <br>LLaVA             | <p>The image features a red bicycle parked in a room, possibly a bedroom. The bicycle is leaning against a bed, which is covered with blankets and sheets. The room appears to be cluttered, with various items scattered around. There are several sports balls in the room, with one located near the left side of the bed, another in the middle, and a third one on the right side. Additionally, there are <b>two bottles</b>, one near the top left corner and the other on the right side of the room. A <b>chair</b> can be seen in the background, and a <b>cup</b> is placed on a surface in the middle of the room.</p> |                                                                                                      |
| <br>LLaVA<br>+ Ours | <p>The image features a red bicycle parked next to a bed. The bicycle is leaning against the bed, and it appears to be an old-fashioned model. The bed is covered with a blanket, and there are several sports balls scattered around the bed. The sports balls are of various sizes and are placed in different positions on the bed. The scene gives off a casual and relaxed atmosphere.</p>                                                                                                                                                                                                                                    |                                                                                                      |
| <br>USER            | Please describe the image in detail.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | <br>Input image |
| <br>LLaVA           | <p>The image features a small, white, square toilet with a hole in the middle, located on a tiled floor. The toilet is positioned in the middle of the room, surrounded by tiles. The floor appears to be dirty, and there is a person standing next to the toilet, possibly using it. In the scene, there are also <b>two bottles</b>, one located near the top left corner and the other near the top right corner of the image. Additionally, there is a <b>cup</b> placed on the floor, closer to the center of the scene.</p>                                                                                                 |                                                                                                      |
| <br>LLaVA<br>+ Ours | <p>The image features a dirty, white toilet with a hole in the middle, located in a bathroom. The toilet is surrounded by tiled walls, giving it a somewhat old and worn appearance. The toilet is situated in the corner of the bathroom, and the tiles are visible on the walls and floor. The overall scene gives off a sense of disrepair and neglect.</p>                                                                                                                                                                                                                                                                     |                                                                                                      |
| <br>USER            | Please describe the image in detail.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | <br>Input image |
| <br>LLaVA           | <p>The image features a man sitting on a couch in a living room, holding a Wii remote in his hand. He appears to be playing a video game, possibly on a Nintendo Wii console. The couch is positioned against a wall, and there is a chair located near the left side of the room. In the background, there are several books scattered around, possibly indicating a reading or relaxation area. A dining table can be seen in the room, with a <b>cup</b> placed on it. Additionally, there is a <b>cell phone</b> resting on the table, and a <b>bottle</b> is located near the edge of the table.</p>                          |                                                                                                      |
| <br>LLaVA<br>+ Ours | <p>The image features a man sitting on a couch, holding a Wii remote in his hand. He is wearing a gray shirt and appears to be enjoying some leisure time. The couch is positioned in a living room, and there are several books scattered around the area. Some books are placed on a shelf. A chair is also visible in the room, situated far away from the couch.</p>                                                                                                                                                                                                                                                           |                                                                                                      |

Figure A5: **Additional qualitative results of our method on LLaVA-1.5-7B with greedy decoding.** On the CHAIR dataset, our method effectively reduces object hallucinations, preventing the model from generating non-existent objects in the image. Hallucinated object words are highlighted in red.

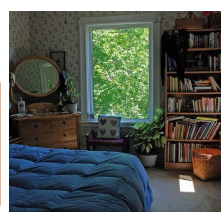
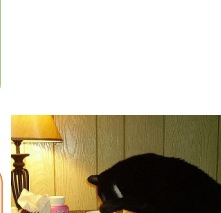
|                                                                                                         |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |                                                                                                      |
|---------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------|
| <br>USER               | Please describe the image in detail.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | <br>Input image   |
| <br>LLaVA              | The image depicts a small bathroom with a white toilet positioned in the center of the room. Next to the toilet, there is a sink with a <b>mirror</b> above it. Above the sink, there is a soap dispenser, and a towel rack is also present in the bathroom. In addition to the main fixtures, there are <b>two cups</b> placed on the floor, one near the left side of the toilet and the other closer to the right side. A <b>bottle</b> can also be seen on the left side of the toilet. The bathroom appears to be clean and well-maintained.                                                                                                                                                                                                                                             |                                                                                                      |
| <br>LLaVA<br>+ OPERA   | The image depicts a small bathroom with a white toilet and a sink. The toilet is positioned towards the left side of the bathroom, while the sink is located on the right side. The bathroom appears to be tiled, giving it a clean and well-maintained appearance. Additionally, there is a toilet brush placed near the toilet, ready for use.                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                      |
| <br>USER               | Please describe the image in detail.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | <br>Input image   |
| <br>LLaVA              | The image features a woman sitting on the edge of a bed in a bedroom. She is wearing a <b>black shirt</b> and appears to be enjoying her time in the room. The bed occupies a significant portion of the scene, extending from the left to the right side of the image. There are <b>two couches</b> in the room, one located on the left side of the bed and the other on the right side. A <b>potted plant</b> can be seen on the left side of the room, adding a touch of greenery to the space. Additionally, there is a <b>cup</b> placed near the right side of the bed, possibly belonging to the woman or someone else in the room.                                                                                                                                                   |                                                                                                      |
| <br>LLaVA<br>+ OPERA   | The image features a woman sitting on a bed in a bedroom. She is positioned near the center of the bed, which is covered with a white comforter. The bed is adorned with a colorful quilt, adding a vibrant touch to the room. The woman appears to be relaxed and enjoying her time in the bedroom.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |                                                                                                      |
| <br>USER               | Please describe the image in detail.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | <br>Input image  |
| <br>LLaVA             | The image features a cozy bedroom with a neatly made bed situated in the center of the room. The bed is adorned with a blue comforter and a blue blanket, creating a warm and inviting atmosphere. In addition to the bed, there is a large window in the room, allowing natural light to fill the space. A potted plant is placed near the window, adding a touch of greenery and life to the room. The room also contains a bookshelf filled with numerous books, showcasing a variety of reading materials for the occupant. A <b>chair</b> is positioned near the bed, providing a comfortable seating area for relaxation or reading. Other items in the room include a <b>vase</b> , a <b>bottle</b> , and a <b>clock</b> , adding to the overall decor and functionality of the space. |                                                                                                      |
| <br>LLaVA<br>+ OPERA | The image depicts a cozy bedroom scene with a bed positioned under a large window. The bed is covered with a blue comforter, and there are several books scattered around the room. A potted plant is placed near the window, adding a touch of greenery to the space. In addition to the bed, there is a dresser and a chair in the room. The dresser is located on the left side of the bed. A mirror is also present in the room, positioned on the left side of the bed. Overall, the room appears to be a comfortable and inviting space for relaxation and reading.                                                                                                                                                                                                                     |                                                                                                      |
| <br>USER             | Please describe the image in detail.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | <br>Input image |
| <br>LLaVA            | The image features a black cat standing on a bed, curiously looking at a cup placed on the bed. The cat appears to be sniffing the cup, possibly intrigued by its contents. In the room, there is a dining table with a few items on it. A bottle can be seen on the table, as well as a <b>bowl</b> and a <b>spoon</b> . Additionally, there are <b>two books</b> placed on the table, one closer to the left side and the other near the center. A remote control is also present on the table, slightly to the left of the center.                                                                                                                                                                                                                                                         |                                                                                                      |
| <br>LLaVA<br>+ OPERA | The image features a black cat standing on a bed, with its head inside a cup. The cat appears to be drinking from the cup, which is placed near the bed. The cat is positioned in the middle of the bed, and the cup is located towards the left side of the bed. The scene captures the cat's curiosity and playfulness as it interacts with the cup.                                                                                                                                                                                                                                                                                                                                                                                                                                        |                                                                                                      |

Figure A6: **Additional qualitative results of our method on LLaVA-1.5-7B with OPERA.** On the CHAIR dataset, our method effectively reduces object hallucinations, preventing the model from generating non-existent objects in the image. Hallucinated object words are highlighted in red.



Figure A7: **Additional qualitative results of our method on LLaVA-1.5-7B with VCD.** On the CHAIR dataset, our method effectively reduces object hallucinations, preventing the model from generating non-existent objects in the image. Hallucinated object words are highlighted in red.





Figure A8: **Additional qualitative results of our method on LLaVA-1.5-7B with PAI.** On the CHAIR dataset, our method effectively reduces object hallucinations, preventing the model from generating non-existent objects in the image. Hallucinated object words are highlighted in red.



Figure A9: **Additional qualitative results of our method on LLaVA-1.5-7B with Devils.** On the CHAIR dataset, our method effectively reduces object hallucinations, preventing the model from generating non-existent objects in the image. Hallucinated object words are highlighted in red.

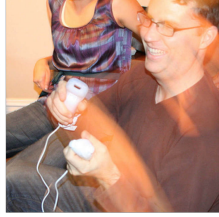
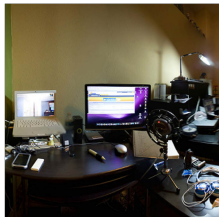
|                                                                                                                                                                                                                                                                                                            |                                                                                                                                                                                                                                                                                     |                                                                                                          |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------|
| <p>USER: Is there a dining table in the image?</p> <p>LLaVA: Yes, there is a <b>dining table</b> in the image, and a man is sitting at it while playing a video game.</p> <p>LLaVA + Ours: No, there is <b>no dining table</b> in the image. The focus is on the man holding a wii remote and smiling.</p> | <p>USER: Is there a tv in the image?</p> <p>LLaVA: Yes, there is a <b>tv</b> in the image, and the man is playing a video game on it.</p> <p>LLaVA + Ours: No, there is <b>no tv</b> in the image. The image shows a man holding a wii remote and smiling.</p>                      |  <p>Input image</p>   |
| <p>USER: Is there a backpack in the image?</p> <p>LLaVA: Yes, there is a <b>backpack</b> in the image, and it is located near the red double-decker bus.</p> <p>LLaVA + Ours: No, there is <b>no backpack</b> in the image.</p>                                                                            | <p>USER: Is there a motorcycle in the image?</p> <p>LLaVA: Yes, there is a <b>motorcycle</b> in the image, and it is parked next to the red double-decker bus.</p> <p>LLaVA + Ours: No, there is <b>no motorcycle</b> in the image. The image features a red double-decker bus.</p> |  <p>Input image</p>   |
| <p>USER: Is there a handbag in the image?</p> <p>LLaVA: Yes, there is a <b>handbag</b> in the image, and it is placed on the ground near the person sitting on the sidewalk.</p> <p>LLaVA + Ours: No, there is <b>no handbag</b> in the image.</p>                                                         | <p>USER: Is there a skateboard in the image?</p> <p>LLaVA: Yes, there is a <b>skateboard</b> in the image, and it is leaning against the yellow shack.</p> <p>LLaVA + Ours: No, there is <b>no skateboard</b> in the image.</p>                                                     |  <p>Input image</p> |
| <p>USER: Is there a bottle in the image?</p> <p>LLaVA: Yes, there is a <b>bottle</b> in the image, which is located on the desk.</p> <p>LLaVA + Ours: No, there is <b>no bottle</b> in the image. The image features a desk with a laptop, a mouse, and a monitor.</p>                                     | <p>USER: Is there a cup in the image?</p> <p>LLaVA: Yes, there is a <b>cup</b> in the image, placed on the desk.</p> <p>LLaVA + Ours: No, there is <b>no cup</b> in the image. The image features a desk with a laptop, a monitor, and a mouse.</p>                                 |  <p>Input image</p> |

Figure A10: **Additional qualitative results of our method on LLaVA-1.5-7B with greedy decoding.** On the POPE dataset, our method correctly identifies objects present in the image. Correct and incorrect answers are highlighted in blue and red, respectively.



Figure A11: **Additional qualitative results of our method on LLaVA-1.5-7B with OPERA.** On the POPE dataset, our method correctly identifies objects present in the image. Correct and incorrect answers are highlighted in blue and red, respectively.



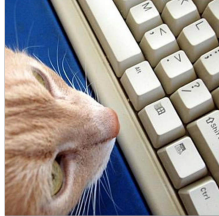
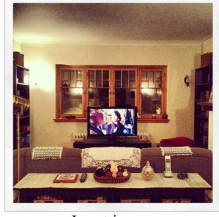
|                                                                                                                                                                                                                                                                                                                                                                     |                                                                                                                                                                                                                                                                                                                   |                                                                                                          |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------|
| <p>USER: Is there a dining table in the image?</p> <p>LLaVA + VCD: <b>Yes</b>, there is a <b>dining table</b> in the image, along with surfboards, backpacks, and people standing at the bus stop.</p> <p>LLaVA + VCD + Ours: <b>No</b>, there is <b>no dining table</b> in the image. The image features a public transmit bus with a man standing next to it.</p> | <p>USER: Is there a cup in the image?</p> <p>LLaVA + VCD: <b>Yes</b>, there is a <b>cup</b> in the image, held by one of the people getting off the bus.</p> <p>LLaVA + VCD + Ours: <b>No</b>, there is <b>no cup</b> in the image.</p>                                                                           |  <p>Input image</p>   |
| <p>USER: Is there a dining table in the image?</p> <p>LLaVA + VCD: <b>Yes</b>, there is a <b>dining table</b> in the image with a computer keyboard on it.</p> <p>LLaVA + VCD + Ours: <b>No</b>, there is <b>no dining table</b> in the image. The image shows a keyboard and a cat sitting next to it.</p>                                                         | <p>USER: Is there a couch in the image?</p> <p>LLaVA + VCD: <b>Yes</b>, there is a <b>couch</b> in the image, with the cat sitting on top of it next to the computer keyboard.</p> <p>LLaVA + VCD + Ours: <b>No</b>, there is <b>no couch</b> in the image. The image features a computer keyboard and a cat.</p> |  <p>Input image</p>   |
| <p>USER: Is there a car in the image?</p> <p>LLaVA + VCD: <b>Yes</b>, there is a <b>car</b> in the image, parked in the background of the living room.</p> <p>LLaVA + VCD + Ours: <b>No</b>, there is <b>no car</b> in the image. It features a living room with a flat screen tv and a window.</p>                                                                 | <p>USER: Is there a cup in the image?</p> <p>LLaVA + VCD: <b>Yes</b>, there is a <b>cup</b> in the image. It is placed on top of a coffee table in a living room.</p> <p>LLaVA + VCD + Ours: <b>No</b>, there is <b>no cup</b> in the image.</p>                                                                  |  <p>Input image</p> |
| <p>USER: Is there a backpack in the image?</p> <p>LLaVA + VCD: <b>Yes</b>, there is a <b>backpack</b> in the image, placed near the table where the cake and coffee are served.</p> <p>LLaVA + VCD + Ours: <b>No</b>, there is <b>no backpack</b> in the image. The focus is on the man and the food items on the table.</p>                                        | <p>USER: Is there a handbag in the image?</p> <p>LLaVA + VCD: <b>Yes</b>, there is a <b>handbag</b> in the image, located near the table where the man is cutting the cake.</p> <p>LLaVA + VCD + Ours: <b>No</b>, there is <b>no handbag</b> in the image.</p>                                                    |  <p>Input image</p> |

Figure A12: **Additional qualitative results of our method on LLaVA-1.5-7B with VCD.** On the POPE dataset, our method correctly identifies objects present in the image. Correct and incorrect answers are highlighted in blue and red, respectively.



Figure A13: **Additional qualitative results of our method on LLaVA-1.5-7B with PAI.** On the POPE dataset, our method correctly identifies objects present in the image. Correct and incorrect answers are highlighted in blue and red, respectively.

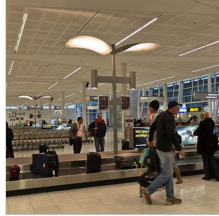
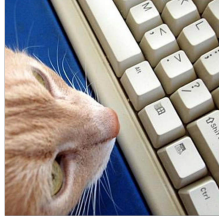
|                                                                                                                                                                                                                                                                                                                      |                                                                                                                                                                                                                                                                                                                  |                                                                                                          |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------|
| <p>USER: Is there a skis in the image?</p> <p>LLaVA + Devils: <b>Yes</b>, there is a <b>skis</b> display in the image, which is located in the airport.</p> <p>LLaVA + Devils + Ours: <b>No</b>, there are <b>no skis</b> in the image. The image shows people walking through an airport with their luggage.</p>    | <p>USER: Is there a car in the image?</p> <p>LLaVA + Devils: <b>Yes</b>, there is a <b>car</b> in the image, which is parked in the airport parking lot.</p> <p>LLaVA + Devils + Ours: <b>No</b>, there is <b>no car</b> in the image.</p>                                                                       |  <p>Input image</p>   |
| <p>USER: Is there a dining table in the image?</p> <p>LLaVA + Devils: <b>Yes</b>, there is a <b>dining table</b> in the image, and the cat is sitting on top of it.</p> <p>LLaVA + Devils + Ours: <b>No</b>, there is <b>no dining table</b> in the image. The image features a cat next to a computer keyboard.</p> | <p>USER: Is there a couch in the image?</p> <p>LLaVA + Devils: <b>Yes</b>, there is a <b>couch</b> in the image, and the cat is sitting on top of it.</p> <p>LLaVA + Devils + Ours: <b>No</b>, there is <b>no couch</b> in the image. The image features a cat next to a computer keyboard.</p>                  |  <p>Input image</p>   |
| <p>USER: Is there a handbag in the image?</p> <p>LLaVA + Devils: <b>Yes</b>, there is a <b>handbag</b> in the image, and the child is holding it.</p> <p>LLaVA + Devils + Ours: <b>No</b>, there is <b>no handbag</b> in the image. The image shows a child in a car seat holding a toy.</p>                         | <p>USER: Is there a truck in the image?</p> <p>LLaVA + Devils: <b>Yes</b>, there is a <b>truck</b> in the image, and the child is sitting in the back of the truck.</p> <p>LLaVA + Devils + Ours: <b>No</b>, there is <b>no truck</b> in the image. The image features a child in a car seat, holding a toy.</p> |  <p>Input image</p> |
| <p>USER: Is there a dining table in the image?</p> <p>LLaVA + Devils: <b>Yes</b>, there is a <b>dining table</b> in the image, which is located next to the bed.</p> <p>LLaVA + Devils + Ours: <b>No</b>, there is <b>no dining table</b> in the image. The image shows a bed with a colorful quilt on it.</p>       | <p>USER: Is there a tv in the image?</p> <p>LLaVA + Devils: <b>Yes</b>, there is a <b>tv</b> in the image, which is located on the right side of the bed.</p> <p>LLaVA + Devils + Ours: <b>No</b>, there is <b>no tv</b> in the image.</p>                                                                       |  <p>Input image</p> |

Figure A14: **Additional qualitative results of our method on LLaVA-1.5-7B with Devils.** On the POPE dataset, our method correctly identifies objects present in the image. Correct and incorrect answers are highlighted in blue and red, respectively.



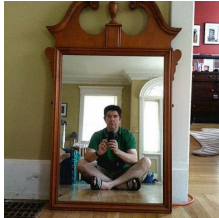
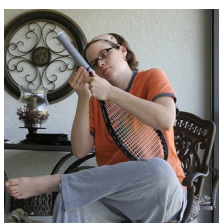
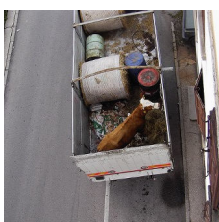
|                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |                                                                                                      |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------|
| <br>USER<br><br>Shikra     | <p>Please describe the image in detail.</p> <p>The image features a man sitting on the floor in front of a mirror, taking a picture of himself. He is wearing a green shirt and appears to be comfortable as he captures the moment. The man is using a cell phone to take the picture, and there is a <b>handbag</b> nearby on the floor. In the room, a <b>couch</b> can be seen in the background, occupying a significant portion of the space. There are also two <b>vases</b> placed at different spots in the room, adding decorative elements to the scene.</p>                                                                                                                                                         | <br>Input image   |
| <br>Shikra<br>+ Ours                                                                                        | <p>The image features a man sitting on the floor in front of a mirror. He is taking a picture of himself using a cell phone, capturing the reflection in the mirror. The man is wearing a green shirt and shorts, which can be seen clearly in the mirror. The primary focus is the man taking a picture of himself through the mirror.</p>                                                                                                                                                                                                                                                                                                                                                                                     |                                                                                                      |
| <br>USER<br><br>Shikra     | <p>Please describe the image in detail.</p> <p>The image shows a man standing in a grassy field, swinging a <b>baseball bat</b> with a <b>baseball glove</b> on the other hand. He appears to be enjoying his time outdoors, actively engaged in a game of baseball. The field is illuminated by a light in the background, creating a warm and inviting atmosphere for the game. Around the field, there are <b>a few cars</b> parked at a distance, suggesting that the field is located in a public area where people can gather and play sports.</p>                                                                                                                                                                        | <br>Input image   |
| <br>Shikra<br>+ Ours                                                                                        | <p>The image features a young man playing frisbee on a grassy field. He is in a throwing motion, holding the frisbee with his right hand while his left hand is positioned behind him. There are several other people in the background, possibly enjoying the outdoors or participating in the game. The scene appears to be set during the evening or night, with the grassy field being illuminated, creating a lively atmosphere.</p>                                                                                                                                                                                                                                                                                       |                                                                                                      |
| <br>USER<br><br>Shikra     | <p>Please describe the image in detail.</p> <p>The image features a woman sitting on a chair with a tennis racket in her hand. She is in the process of looking at a <b>cell phone</b>, which she is holding in her other hand. The chair is positioned in a way that allows her to comfortably check her <b>phone</b> while relaxing. Additionally, there are a few <b>books</b> placed nearby, suggesting that the woman might be in a casual and comfortable setting, possibly at home or a public space with a <b>bookshelf</b>. The scene captures a moment of leisure or relaxation for the woman.</p>                                                                                                                    | <br>Input image  |
| <br>Shikra<br>+ Ours                                                                                       | <p>The image features a woman sitting in a chair with a tennis racket in her hand. She appears to be wearing glasses and is likely dressed in a casual outfit. The woman is holding the tennis racket up in front of her, possibly in anticipation of playing or warming up for a game.</p>                                                                                                                                                                                                                                                                                                                                                                                                                                     |                                                                                                      |
| <br>USER<br><br>Shikra | <p>Please describe the image in detail.</p> <p>The image features an aerial view of a large truck driving down a street. The truck is filled with various items, including a cow, a <b>trailer</b> full of <b>containers</b>, and numerous garbage-filled barrels. The cow can be seen occupying a significant portion of the back of the truck, while the barrels are scattered throughout the truck bed. A <b>person</b> is visible near the truck, possibly observing the scene or involved in the transportation process. The image captures the essence of a busy street with a truck carrying a diverse load.</p>                                                                                                         | <br>Input image |
| <br>Shikra<br>+ Ours                                                                                      | <p>The image depicts a large truck driving down a street, its cargo compartment filled with various items. In the back of the truck, a cow can be seen, likely being transported to a farm or another location. The truck's side displays a mix of different materials, including two large reels of rope and a crushed can. Additionally, there is a blue barrel on the truck, which contributes to the diverse load it carries. The truck appears to be moving at a moderate speed, and its cargo compartment is quite full, showcasing the diverse items it carries.</p>                                                                                                                                                     |                                                                                                      |
| <br>USER<br><br>Shikra | <p>Please describe the image in detail.</p> <p>The image features a group of men in uniform, working together in a navy ship's control room. They are wearing headphones and sunglasses, focusing on their tasks as they operate various communication equipment. There are at least five individuals in the room, with some sitting at a table and others standing near the windows. A laptop is placed on the table in front of the men, and one of the individuals is using a handheld radio on his shoulder. The room is equipped with a <b>TV</b>, which is visible in the background. Additionally, a <b>book</b> can be seen on the table, possibly a reference material or guide for the tasks they are performing.</p> | <br>Input image |
| <br>Shikra<br>+ Ours                                                                                      | <p>The image features a group of men in navy uniforms, working together in a control room. They are using headphones and a microphone to communicate with each other, with some of them also wearing sunglasses. There is a laptop in the scene, one placed on the left side of the room. The men are working diligently, likely engaged in their tasks, as they navigate the control room and communicate with one another.</p>                                                                                                                                                                                                                                                                                                |                                                                                                      |

Figure A15: **Additional qualitative results of our method on Shikra-7B with greedy decoding.** On the CHAIR dataset, our method effectively reduces object hallucinations, preventing the model from generating non-existent objects in the image. Hallucinated object words are highlighted in red.

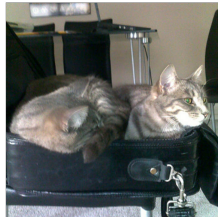
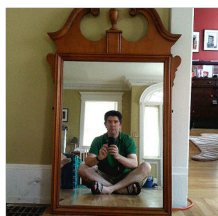
|                                                                                                                    |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |                                                                                                      |
|--------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------|
| <br>USER                          | Please describe the image in detail.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | <br>Input image   |
| <br>Shikra<br>+ OPERA             | The image depicts a cozy scene with two grey cats lying down inside a black suitcase. One cat is positioned towards the left side of the suitcase, while the other cat is on the right side. They appear to be relaxed and comfortable in their shared travel luggage. In addition to the cats and the suitcase, there are a few other objects in the scene. A <b>laptop</b> can be spotted on the left side of the image, while a chair is situated nearer on the right side. A <b>dining table</b> and a tabletop are also visible in the scene. |                                                                                                      |
| <br>Shikra<br>+ OPERA<br>+ Ours   | The image depicts a cozy scene with two cats sitting together inside a black suitcase. One cat can be seen on the left side of the suitcase, while the other cat is on the right side. They appear to be relaxing and enjoying each other's company in the luggage. In the background, a chair is partially visible, situated behind the suitcase where the cats are sitting.                                                                                                                                                                      |                                                                                                      |
| <br>USER                          | Please describe the image in detail.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | <br>Input image   |
| <br>Shikra<br>+ OPERA             | The image features a man sitting on the floor in front of a mirror, taking a picture of himself with a <b>digital camera</b> . He is wearing a green shirt and appears to be comfortable as he captures the moment with the camera. The room has a cozy atmosphere with a <b>couch</b> in the background to the right and <b>two chairs</b> nearby. A <b>dog</b> can be seen resting or relaxing on the floor, located near the man's right hand side at the lower-height perspective of rug on a wooden floor.                                    |                                                                                                      |
| <br>Shikra<br>+ OPERA<br>+ Ours   | The image features a man sitting on the floor in front of a mirror. He is taking a picture of himself using a cell phone, capturing the reflection in the mirror. The man is wearing a green shirt and shorts, giving a casual and relaxed vibe to the scene.                                                                                                                                                                                                                                                                                      |                                                                                                      |
| <br>USER                          | Please describe the image in detail.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | <br>Input image  |
| <br>Shikra<br>+ OPERA             | The image features a white building with a prominent clock on the front of it. The clock is positioned near the center of the building, making it easily visible to passersby. The building's white facades give it a clean and modern appearance. In the vicinity of the building, there is a <b>motorcycle</b> parked on the right side, as well as a <b>bicycle</b> on the left side of the scene. These additional objects add to the overall atmosphere of the area.                                                                          |                                                                                                      |
| <br>Shikra<br>+ OPERA<br>+ Ours | The image features a white building with a large clock mounted on the front of it. The clock is positioned near the center of the building, making it easily visible to passersby. The building's white color adds a sense of cleanliness and simplicity to the scene.                                                                                                                                                                                                                                                                             |                                                                                                      |
| <br>USER                        | Please describe the image in detail.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | <br>Input image |
| <br>Shikra<br>+ OPERA           | The image depicts a peaceful scene at a viewpoint overlooking a large river flowing through a valley. A single wooden bench is placed at the edge of the hill, providing a comfortable seating area for visitors to relax and take in the scenic surroundings. In the background, a bridge can be seen crossing the river, adding to the picturesque setting. Near the bench, a <b>handbag</b> and a <b>backpack</b> are placed, likely belonging to visitors enjoying a walk or hike in the area.                                                 |                                                                                                      |
| <br>Shikra<br>+ OPERA<br>+ Ours | The image features a wooden bench situated at the top of a hill, overlooking a beautiful landscape that includes a river and a bridge. The bench is facing the scenic view, offering a picturesque spot to sit and enjoy the view.                                                                                                                                                                                                                                                                                                                 |                                                                                                      |
| <br>USER                        | Please describe the image in detail.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | <br>Input image |
| <br>Shikra<br>+ OPERA           | The image features a large bathroom with a separate shower and bathtub. The main focus is on a well-lit sink with a mirror above it, which is accompanied by a vanity. On the countertop, there is a <b>bottle of wine</b> placed near the sink. In addition to a bathtub with a red shower curtain, the bathroom also includes a <b>toilet</b> with a matching red bath mat. The overall color scheme of the bathroom is predominantly white, giving it a clean and elegant appearance.                                                           |                                                                                                      |
| <br>Shikra<br>+ OPERA<br>+ Ours | The image features a clean bathroom with a sink and a bathtub. The sink has a large mirror above it and is accompanied by a shower curtain. The bathroom appears spotless and well-maintained as a whole.                                                                                                                                                                                                                                                                                                                                          |                                                                                                      |

Figure A16: **Additional qualitative results of our method on Shikra-7B with OPERA.** On the CHAIR dataset, our method effectively reduces object hallucinations, preventing the model from generating non-existent objects in the image. Hallucinated object words are highlighted in red.

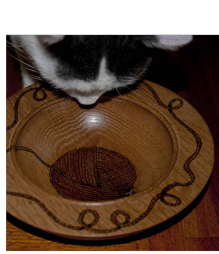
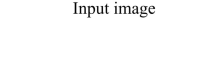
|                                                                                                                  |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |                                                                                       |
|------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|
| <br>USER                        | Please describe the image in detail.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |    |
| <br>Shikra<br>+ VCD             | The image depicts a scenic view of a river flowing through a valley, with a wooden bench situated on a grassy hill overlooking the surrounding landscape. The bench is located at a pleasant viewpoint, offering a great vantage point of the river and the hills beyond. In the scene, there is a <b>boat</b> visible near the river, likely enjoying the calm waters as it sails along the valley. Additionally, a <b>train</b> can be seen on the tracks, following the path of the river, adding another element of interest to the picturesque landscape.                                                                                                       | Input image                                                                           |
| <br>Shikra<br>+ VCD<br>+ Ours   | The scene features a wooden bench sitting at the edge of a grassy hill, overlooking a scenic view that includes a river running through a valley. The bench is facing the water, offering a picturesque and relaxing spot to sit and take in the surroundings.                                                                                                                                                                                                                                                                                                                                                                                                       |    |
| <br>USER                        | Please describe the image in detail.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 | Input image                                                                           |
| <br>Shikra<br>+ VCD             | The image depicts a serene outdoor scene featuring a lone pine tree growing next to a wooden bench. The bench is positioned in a clearing beneath the tree, providing a resting spot for visitors or passers-by. The area seems to be a quiet, picturesque location, possibly overlooking a beautiful landscape, such as a valley or town below. There are <b>a few people</b> scattered around the scene, likely enjoying the tranquility of the location. Some are close to the bench, while others are closer to the tree or standing at a slight distance. A <b>handbag</b> can be seen on the ground near the bench, possibly belonging to one of the visitors. |    |
| <br>Shikra<br>+ VCD<br>+ Ours   | The image features a beautiful pine tree with a wooden bench positioned right underneath it. The tree is tall and slender, reaching towards the sky. It appears to be a sunny day, casting light on the tree and bench, making them stand out in the scene. The bench is small but close to the tree, providing a cozy spot to sit and enjoy the view. The landscape in the background consists of a mountain, which can be seen peering over the horizon. There is also a patch of grass next to the tree and bench.                                                                                                                                                |   |
| <br>USER                       | Please describe the image in detail.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 | Input image                                                                           |
| <br>Shikra<br>+ VCD           | The image captures a street scene with a one-way sign pointing to the right. A wall with graffiti is visible in the background, adding an artistic touch to the urban landscape. A spray paint can be seen as part of the graffiti, highlighting the creativity present in the area. In the foreground, there is a bench providing a place for people to sit and observe the surroundings. <b>Several cars</b> are parked nearby, with two located on the right side of the scene and one closer to the center. A <b>fire hydrant</b> is also present close to the center of the image.                                                                              |  |
| <br>Shikra<br>+ VCD<br>+ Ours | The image features a one-way street sign positioned next to a white wall. On the wall, there is a graffiti drawing of a hand, adding a unique decorative touch to the signage. In the background, a bench can be seen near the left side of the scene.                                                                                                                                                                                                                                                                                                                                                                                                               |  |
| <br>USER                      | Please describe the image in detail.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 | Input image                                                                           |
| <br>Shikra<br>+ VCD           | The image features a black and white cat sitting on a wooden dining table, intently inspecting a wooden bowl. The bowl has a ball of yarn or a skein of string in it, catching the cat's attention. The cat appears to be eating or playing with the <b>ball</b> , as it moves closer and closer to it. A <b>second cat</b> can be seen in the background, away from the yarn-filled bowl, making the scene look like it could be part of a playful cat-oriented environment.                                                                                                                                                                                        |  |
| <br>Shikra<br>+ VCD<br>+ Ours | The image features a black and white cat sitting on a dining table, curiously inspecting a wooden bowl. The cat is the main focus of the scene, occupying a significant portion of the frame. The bowl, which is empty, is placed near the cat, drawing its attention. The table serves as a surface for the cat to explore, as it looks at the bowl with interest.                                                                                                                                                                                                                                                                                                  |  |

Figure A17: **Additional qualitative results of our method on Shikra-7B with VCD.** On the CHAIR dataset, our method effectively reduces object hallucinations, preventing the model from generating non-existent objects in the image. Hallucinated object words are highlighted in red.





USER

Please describe the image in detail.



Shikra + PAI

The image shows a man in a brown shirt, standing in a kitchen and preparing a sandwich on a white countertop. He is in the process of slicing a large sandwich, which appears to be a sub sandwich. There are several items on the countertop, including a bottle, a **cup**, a knife, and a pair of **scissors**. A **microwave** is also visible in the background. The scene depicts a man working on a sandwich in the kitchen, suggesting that he might be preparing a meal or a snack.



Shikra + PAI + Ours

The image shows a man standing in a kitchen, preparing a sandwich by cutting a large sandwich in half. He appears to be making a sandwich with a knife, standing at a kitchen counter. The sandwich is placed on a cutting board, and there are several ingredients visible on the counter, suggesting a variety of ingredients to be used in the sandwich. The man seems to be focused on the sandwich, and he is likely in the process of preparing it.



Input image



USER

Please describe the image in detail.



Shikra + PAI

The image shows two white bowls filled with a creamy soup or chowder, with small portions of broccoli in it. There are **also two large bowls placed on a cutting board**, adding to the scene. In the image, a spoon can be seen resting near the bowls, suggesting that the soup or sauce could be served as a side dish or a light meal. The presence of a **fork** also indicates that the scene is set in a dining area, where people can enjoy their meal.



Shikra + PAI + Ours

The image shows two white bowls filled with a creamy soup containing broccoli and other vegetables. The bowls are placed on a wooden dining table, and there is a spoon resting near them. The dining table is set up in the scene to display the bowls and the soup.



Input image



USER

Please describe the image in detail.



Shikra + PAI

The image shows a white bowl filled with green apples, which are placed in a white bowl. There are several apples in the bowl, including a mix of green and red apples. The bowl is filled with a mix of green apples, with some **red apples** as well. The green apples are arranged in various positions and orientations, creating an interesting composition. The apples are placed in a white bowl, which is placed on a **dining table**.



Shikra + PAI + Ours

The image shows a close-up of a bowl filled with numerous green apples, which are placed in a white bowl. There are various apples in the bowl, with some positioned in the front and others in the back. The apples are arranged in different rows and rows, creating a visually appealing composition. The bowl is filled with several apples, some of which are stacked on top of a smaller apples, creating a vibrant green color palette.



Input image



USER

Please describe the image in detail.



Shikra + PAI

The image shows a park with a display of many colorful umbrellas hanging in the air, creating a vibrant and lively atmosphere. These umbrellas are suspended **upside down**, with their unique positioning creating an artistic and visually appealing scene. The umbrellas are spread throughout the park, with some closer to the ground and others higher up in the trees. There are **several people** in the scene, possibly admiring the display or participating in it. The scene also includes a **bench** located in the middle of the image, providing a place for people to sit and enjoy the view of the colorful umbrellas above them.



Shikra + PAI + Ours

The image shows a park with a display of many colorful umbrellas hanging from trees. These umbrellas are arranged in a forest area, creating a visually appealing and artistic scene. The umbrellas are of various sizes and are spread throughout the forest. The vibrant colors of the umbrellas contrast with the green foliage of the trees in the background, adding a unique and lively touch to the landscape.



Input image

Figure A18: **Additional qualitative results of our method on Shikra-7B with PAI.** On the CHAIR dataset, our method effectively reduces object hallucinations, preventing the model from generating non-existent objects in the image. Hallucinated object words are highlighted in red.



USER

Please describe the image in detail.



Shikra

+ Devils

The image features a wooden dining table with a total of five wine bottles placed on it. Surrounding the table, there are wine glasses and cups, indicating a wine tasting event. The table is filled with numerous wine glasses and bottles, emphasizing the guests' experience. The table is set up as a wine tasting event, with the purpose of providing an enjoyable atmosphere for the attendees. The table has a total of **seven** wine bottles and **six** wine glasses placed on it, showcasing a variety of wines to be tasted. The table also has a **vase** and a **potted plant** decoration nearby, adding a touch of greenery and elegance to the setting.



Shikra

+ Devils



Ours

The image features a wooden dining table with various bottles of wine and wine glasses placed on it. The table is set up like a bar, with multiple wine bottles and wine glasses arranged neatly. The table is surrounded by a collection of wine bottles and wine glasses, creating an atmosphere of relaxation and enjoyment.



Input image



USER

Please describe the image in detail.



Shikra

+ Devils

The image features a man sitting on a park bench surrounded by a large flock of pigeons, with several birds eagerly gathering around him. The man is wearing a baseball cap, and it seems that he is the center of attention in this scene. In the background, there are **multiple cars** and **a few traffic lights** visible, indicating that the scene takes place in an urban setting. There are also **a few cars** and a truck parked nearby, adding to the overall atmosphere of the scene.



Shikra

+ Devils



Ours

The scene features a man sitting on a park bench surrounded by a large flock of pigeons, with several birds taking a liking to his left. The man is wearing a baseball cap, and the birds are gathered around him, with some birds also on the bench next to him. In the image, there are numerous pigeons present, with a mix of birds surrounding the man. The birds are eager to get close to the man, and some are even perched on the bench next to him. The man is likely a professional baseball player, and the birds are likely attracted to him as a result of their presence in the scene.



Input image



USER

Please describe the image in detail.



Shikra

+ Devils

The image showcases two laptops, one smaller and another larger, sitting on a white desk in a room. The smaller laptop is placed in **front** of the larger laptop, while the larger laptop is situated behind it. The two laptops are connected to a **keyboard** and a cell phone, indicating that they might be used for streaming music or other entertainment purposes. In the background, there is a **clock** on the wall, adding a decorative touch to the room. **Four people** can be seen in various positions throughout the image, with two of them closer to the foreground and the other two further away. Additionally, there are **two cell phones** in the scene, one located closer to the larger laptop and the other near the smaller laptop. A clock can be seen in the image. The setup creates an atmosphere of technology and leisure in the room.



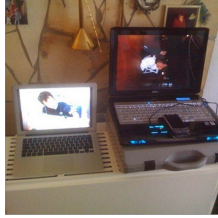
Shikra

+ Devils



Ours

The image features two laptops placed on top of a white desk. One laptop is positioned closer to the left side of the room, while the other laptop is located towards the right side of the room. The wall behind the laptops has a picture hanging on the wall.



Input image



USER

Please describe the image in detail.



Shikra

+ Devils

The image features a beautifully decorated birthday cake with a beach theme. The cake is placed on a **table**, and it has a name on it, "Happy Birthday" is written on the front. The cake is decorated with a combination of lettering and palm trees, which makes it a unique and creative design. The cake is displayed on a **dining table**, and it's available as an additional decoration or centerpiece for a **50th** birthday celebration. The cake is placed on a **dining table**, and it's a great addition to the event. The cake is covered with a blue tablecloth, adding a festive touch to the overall theme.



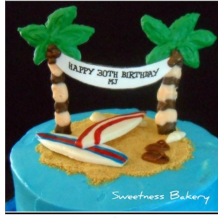
Shikra

+ Devils



Ours

The image features a birthday cake with a beach theme, complete with a surfboard and palm trees. The cake is decorated with a blue background, and the words "Happy Birthday" is written on top.



Input image

Figure A19: **Additional qualitative results of our method on Shikra-7B with Devils.** On the CHAIR dataset, our method effectively reduces object hallucinations, preventing the model from generating non-existent objects in the image. Hallucinated object words are highlighted in red.

|                                                                                                                                                                                                                                                                                                                                   |                                                                                                                                                                                                                                                                                                                              |                                                                                                          |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------|
| <p>USER: Is there a bicycle in the image?</p> <p>Shikra: Yes, there is a <b>bicycle</b> in the image.</p> <p>Shikra + Ours: No, there is <b>no bicycle</b> in the image. The image features a man riding a motorcycle in front of a bus.</p>                                                                                      | <p>USER: Is there a traffic light in the image?</p> <p>Shikra: Yes, there is a <b>traffic light</b> in the image, and a man on a motorcycle is riding in front of a bus.</p> <p>Shikra + Ours: No, there is <b>no traffic light</b> in the image. The photo features a man riding a motorcycle, a green bus, and a road.</p> |  <p>Input image</p>   |
| <p>USER: Is there a dining table in the image?</p> <p>Shikra: Yes, there is a <b>dining table</b> in the image, and it is located in the middle of the bus.</p> <p>Shikra + Ours: No, there is <b>no dining table</b> in the image. The image features a large green and white bus with its door open, parking on the street.</p> | <p>USER: Is there a traffic light in the image?</p> <p>Shikra: Yes, there is a <b>traffic light</b> visible in the image, and the bus stopped at the traffic light.</p> <p>Shikra + Ours: The answer is <b>no</b>.</p>                                                                                                       |  <p>Input image</p>   |
| <p>USER: Is there a cup in the image?</p> <p>Shikra: Yes, there is a <b>bloody knife</b> in the image, and it is in the form of a fruit with a knife in it.</p> <p>Shikra + Ours: No, there is a knife in the image, but there is <b>no cup</b> visible.</p>                                                                      | <p>USER: Is there a person in the image?</p> <p>Shikra: Yes, there is a <b>person</b> in the image who is holding a pair of bloody knives.</p> <p>Shikra + Ours: No, there is a knife in the image, but there is <b>no person</b> visible.</p>                                                                               |  <p>Input image</p> |
| <p>USER: Is there a laptop in the image?</p> <p>Shikra: No, there is <b>no laptop</b> in the image. The image features a large computer setup with two monitors on a desk.</p> <p>Shikra + Ours: Yes, there is a <b>laptop</b> in the image, placed next to the desktop computer on the desk.</p>                                 | <p>USER: Is there a cell phone in the image?</p> <p>Shikra: Yes, there is a <b>cell phone</b> in the image.</p> <p>Shikra + Ours: The answer is <b>no</b>.</p>                                                                                                                                                               |  <p>Input image</p> |

Figure A20: **Additional qualitative results of our method on Shikra-7B with greedy decoding.** On the POPE dataset, our method correctly identifies objects present in the image. Correct and incorrect answers are highlighted in blue and red, respectively.



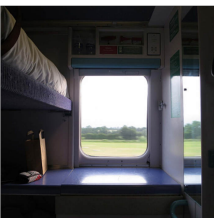
|                                                                                                                                                                                                                                                                    |                                                                                                                                                                                                                                                                |                                                                                                          |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------|
| <p>USER<br/>Is there a person in the image?</p> <p>Shikra<br/>+ OPERA<br/>The answer is <b>no</b>.</p> <p>Shikra<br/>+ OPERA<br/>+ Ours<br/>Yes, there is a <b>person</b> in the image, standing in the water at the beach.</p>                                    | <p>USER<br/>Is there a dining table in the image?</p> <p>Shikra<br/>+ OPERA<br/>Yes, there is a <b>dining table</b> in the image, which is set up on the beach.</p> <p>Shikra<br/>+ OPERA<br/>+ Ours<br/>No, there is <b>no dining table</b> in the image.</p> |  <p>Input image</p>   |
| <p>USER<br/>Is there a book in the image?</p> <p>Shikra<br/>+ OPERA<br/>Yes, there is a <b>book</b> on the bed in the image.</p> <p>Shikra<br/>+ OPERA<br/>+ Ours<br/>The answer is <b>no</b>.</p>                                                                 | <p>USER<br/>Is there a cup in the image?</p> <p>Shikra<br/>+ OPERA<br/>Yes, there is a <b>cup</b> in the image.</p> <p>Shikra<br/>+ OPERA<br/>+ Ours<br/>The answer is <b>no</b>.</p>                                                                          |  <p>Input image</p>   |
| <p>USER<br/>Is there a cup in the image?</p> <p>Shikra<br/>+ OPERA<br/>Yes, there is a bloody knife in the image, and it is hovering over either the top half of an apple or a severed head.</p> <p>Shikra<br/>+ OPERA<br/>+ Ours<br/>The answer is <b>no</b>.</p> | <p>USER<br/>Is there a chair in the image?</p> <p>Shikra<br/>+ OPERA<br/>Yes, there is a <b>chair</b> in the image, and it has a knife in it.</p> <p>Shikra<br/>+ OPERA<br/>+ Ours<br/>The answer is <b>no</b>.</p>                                            |  <p>Input image</p> |
| <p>USER<br/>Is there a bed in the image?</p> <p>Shikra<br/>+ OPERA<br/>(No answer)</p> <p>Shikra<br/>+ OPERA<br/>+ Ours<br/>Yes, there is a <b>bed</b> in the image.</p>                                                                                           | <p>USER<br/>Is there a sports ball in the image?</p> <p>Shikra<br/>+ OPERA<br/>Yes, there is a <b>sports ball</b> in the image.</p> <p>Shikra<br/>+ OPERA<br/>+ Ours<br/>The answer is <b>no</b>.</p>                                                          |  <p>Input image</p> |

Figure A21: **Additional qualitative results of our method on Shikra-7B with OPERA.** On the POPE dataset, our method correctly identifies objects present in the image. Correct and incorrect answers are highlighted in blue and red, respectively.

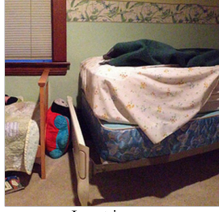
|                                                                                                                                                                                                                                                                            |                                                                                                                                                                                                                                                                                                                                                         |                                                                                                          |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------|
| <p>USER: Is there a couch in the image?</p> <p>Shikra + VCD: <b>Yes</b>, there is a <b>couch</b> in the image.</p> <p>Shikra + VCD + Ours: There are two vases with flowers in the image, <b>not</b> a <b>couch</b>. The vases are placed on a wooden table or shelf.</p>  | <p>USER: Is there a cup in the image?</p> <p>Shikra + VCD: <b>Yes</b>, there is a <b>cup</b> in the image.</p> <p>Shikra + VCD + Ours: <b>No</b>, there are <b>no cups</b> in the image. There are only two flower vases.</p>                                                                                                                           |  <p>Input image</p>   |
| <p>USER: Is there a dining table in the image?</p> <p>Shikra + VCD: <b>Yes</b>, there is a <b>dining table</b> in the image.</p> <p>Shikra + VCD + Ours: The image does <b>not</b> show a <b>dining table</b>. There is a red double-decker bus on a street in a city.</p> | <p>USER: Is there a traffic light in the image?</p> <p>Shikra + VCD: There is a bus in the image, and it is sitting in the street next to a sidewalk. There is also a <b>traffic light</b> visible in the background.</p> <p>Shikra + VCD + Ours: The image does <b>not</b> show a <b>traffic light</b>, so there is no traffic light in the image.</p> |  <p>Input image</p>   |
| <p>USER: Is there a person in the image?</p> <p>Shikra + VCD: <b>Yes</b>, there is a person walking near the traffic light in the image.</p> <p>Shikra + VCD + Ours: The answer is <b>no</b>.</p>                                                                          | <p>USER: Is there a truck in the image?</p> <p>Shikra + VCD: The answer is <b>no</b>.</p> <p>Shikra + VCD + Ours: The answer is <b>yes</b>.</p>                                                                                                                                                                                                         |  <p>Input image</p> |
| <p>USER: Is there a car in the image?</p> <p>Shikra + VCD: <b>Yes</b>, there is a <b>car</b> in the image, and it is a miniature one. It is placed near the bed in the room.</p> <p>Shikra + VCD + Ours: The answer is <b>no</b>.</p>                                      | <p>USER: Is there a bed in the image?</p> <p>Shikra + VCD: <b>(No answer)</b></p> <p>Shikra + VCD + Ours: <b>Yes</b>, there is a <b>bed</b> in the image, and it appears to be situated in a bedroom.</p>                                                                                                                                               |  <p>Input image</p> |

Figure A22: **Additional qualitative results of our method on Shikra-7B with VCD.** On the POPE dataset, our method correctly identifies objects present in the image. Correct and incorrect answers are highlighted in blue and red, respectively.

|                                                                                                                                                                                                                          |                                                                                                                                                                                                                                                              |                                                                                                          |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------|
| <p>USER: Is there a refrigerator in the image?</p> <p>Shikra + PAI: Yes, there is a <b>refrigerated truck</b> in the image.</p> <p>Shikra + PAI + Ours: The answer is <b>no</b>.</p>                                     | <p>USER: Is there a dining table in the image?</p> <p>Shikra + PAI: Yes, there is a <b>dining table</b> in the middle of the street in the image.</p> <p>Shikra + PAI + Ours: No, there is <b>no dining table</b> in the image. The answer is <b>no</b>.</p> |  <p>Input image</p>   |
| <p>USER: Is there a bus in the image?</p> <p>Shikra + PAI: Yes, there is a large group of people entering an airplane.</p> <p>Shikra + PAI + Ours: No, there is <b>no bus</b> in the image. The answer is <b>no</b>.</p> | <p>USER: Is there a potted plant in the image?</p> <p>Shikra + PAI: Yes, there is a large <b>puss</b> in the middle of the picture.</p> <p>Shikra + PAI + Ours: No, there is <b>no potted plant</b> in the image. The answer is <b>no</b>.</p>               |  <p>Input image</p>   |
| <p>USER: Is there a laptop in the image?</p> <p>Shikra + PAI: No, there is <b>no laptop</b> in the image. So the answer is no.</p> <p>Shikra + PAI + Ours: Yes, there is a <b>laptop</b> in the image.</p>               | <p>USER: Is there a cell phone in the image?</p> <p>Shikra + PAI: Yes, there is a keyboard on the desk in the image.</p> <p>Shikra + PAI + Ours: The answer is <b>no</b>.</p>                                                                                |  <p>Input image</p> |
| <p>USER: Is there a dining table in the image?</p> <p>Shikra + PAI: Yes, there is a bus in the image.</p> <p>Shikra + PAI + Ours: The answer is <b>no</b>.</p>                                                           | <p>USER: Is there a chair in the image?</p> <p>Shikra + PAI: The answer is <b>yes</b>.</p> <p>Shikra + PAI + Ours: The answer is <b>no</b>.</p>                                                                                                              |  <p>Input image</p> |

Figure A23: **Additional qualitative results of our method on Shikra-7B with PAI.** On the POPE dataset, our method correctly identifies objects present in the image. Correct and incorrect answers are highlighted in blue and red, respectively.



Figure A24: **Additional qualitative results of our method on Shikra-7B with Devils.** On the POPE dataset, our method correctly identifies objects present in the image. Correct and incorrect answers are highlighted in blue and red, respectively.

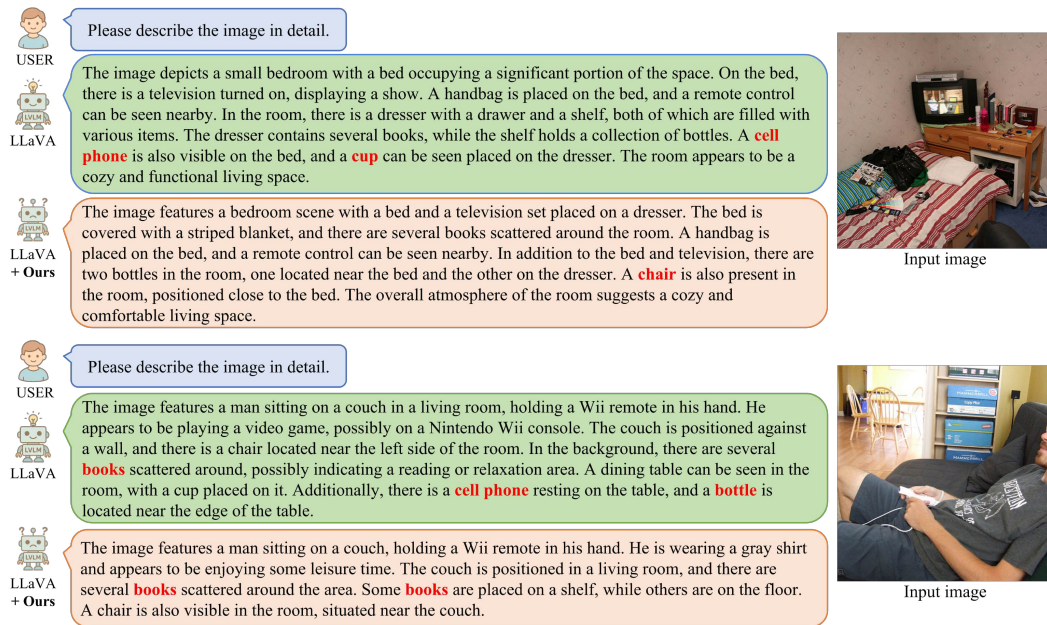


Figure A25: **Failure cases of our method on LLaVA-1.5-7B with greedy decoding.** On the CHAIR dataset, our method effectively reduces object hallucinations but fails to completely prevent the generation of non-existent objects. Hallucinated object words are highlighted in red.