Hallucinatory Image Tokens: A Training-free `EAZY` Approach to Detecting and Mitigating Object Hallucinations in LVLMs

Liwei Che
lc1279@cs.rutgers.edu

Tony Qingze Liu
ql236@scarletmail.rutgers.edu

Jing Jia
jing.jia@rutgers.edu

Weiyi Qin
wq50@cs.rutgers.edu

Ruixiang Tang
ruixiang.tang@rutgers.edu

Vladimir Pavlovic
vladimir@cs.rutgers.edu

Rutgers University
New Brunswick, NJ, USA 08901

Abstract

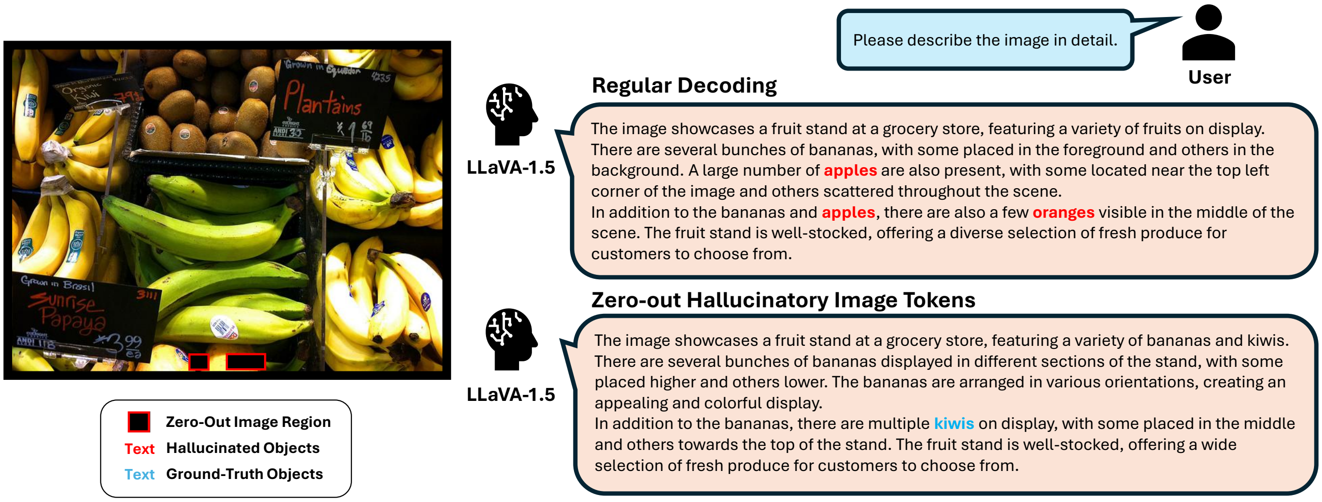
Despite their remarkable potential, Large Vision-Language Models (LVLMs) still face challenges with object hallucination, a problem where their generated outputs mistakenly incorporate objects that do not actually exist. Although most works focus on addressing this issue within the language-model backbone, our work shifts the focus to the image input source, investigating how specific image tokens contribute to hallucinations. Our analysis reveals a striking finding: a small subset of image tokens with high attention scores are the primary drivers of object hallucination. By removing these hallucinatory image tokens (only 1.5% of all image tokens), the issue can be effectively mitigated. This finding holds consistently across different models and datasets. Building on this insight, we introduce `EAZY`*, a novel, training-free method that automatically identifies and* ***E****liminate h****A****llucinations by* ***Z****eroing out hallucinator****Y*** *image tokens. We utilize* `EAZY` *for unsupervised object hallucination detection, achieving 15% improvement compared to previous methods. Additionally,* `EAZY` *demonstrates remarkable effectiveness in mitigating hallucinations while preserving model utility and seamlessly adapting to various LVLM architectures.*

1. Introduction

Large Vision-Language Models (LVLMs) [4, 18, 20] have achieved remarkable advancements, seamlessly integrating visual recognition and language understanding to produce outputs that are both coherent and contextually aligned. Despite these successes, LVLMs face a critical challenge from hallucination [21]. Hallucination in LVLMs manifests as generating incorrect responses that either contradict the provided image input or deviate from user instructions. This problem not only compromises LVLMs reliability but also limits their broader applicability in real-world scenarios. Among the various forms of hallucination, object hallucination (OH) [28] is particularly prevalent. This occurs when LVLMs either misidentify objects or incorrectly perceive hallucinatory objects (HO) as being present in an image. Addressing this issue is essential to enhance the accuracy and trustworthiness of LVLMs for practical applications.

Various studies [9, 28, 29] have been conducted to investigate the underlying causes of OHs in LVLMs. A significant portion of existing research attributes this issue to the biases from the language part. In [14, 16, 36], the authors argue that statistical biases in the textual training data are a major contributing factor. These biases include the long-tailed distribution of object occurrence frequencies [16] and the co-occurrence relationships between different objects [28, 36]. Some other works [13, 14, 31] believe that prior knowledge from the powerful language model overrides input information, resulting in various hallucinations. Consequently, these studies employ contrastive decoding [15], which achieves decoding correction by comparing token probability distributions before and after modification to image input. Another line of research noticed that the language model pays less attention to the image tokens but more to the most recent text token during the generation process therefore applying generation backtracking [10, 32]. While these pioneering efforts have provided valuable insights into the causes of visual hallucinations, a key question remains: *How does visual input influence the generation of visual hallucinations?* This study aims to explore this critical aspect in depth to better understand hallucination phenomena from the vision perspective.

To address this gap, we investigated how LVLMs process visual inputs and how this contributes to visual OHs. Although using the distribution of attention weights to in-



Please describe the image in detail.

The image showcases a fruit stand at a grocery store, featuring a variety of fruits on display. There are several bunches of bananas, with some placed in the foreground and others in the background. A large number of **apples** are also present, with some located near the top left corner of the image and others scattered throughout the scene.

In addition to the bananas and **apples**, there are also a few **oranges** visible in the middle of the scene. The fruit stand is well-stocked, offering a diverse selection of fresh produce for customers to choose from.

The image showcases a fruit stand at a grocery store, featuring a variety of bananas and kiwis. There are several bunches of bananas displayed in different sections of the stand, with some placed higher and others lower. The bananas are arranged in various orientations, creating an appealing and colorful display.

In addition to the bananas, there are multiple **kiwis** on display, with some placed in the middle and others towards the top of the stand. The fruit stand is well-stocked, offering a wide selection of fresh produce for customers to choose from.

Figure 1. Removing three image tokens results in the elimination of the hallucinated objects, "apples" and "oranges", and reveals the real object "kiwis".

terpret visual token contributions is a straightforward approach [8, 24, 33], many layers exhibit noisy attention patterns with no clear correlation between generated object-related tokens and specific image regions. Through a comprehensive analysis of different layers, as shown in Figure 3, we found that LLaVA and similar models predominantly extract object information in their middle-to-late layers. For example, in LLaVA-1.5 (7b), a 32-layer transformer model, we reveal that in the middle layers, such as Layer 15, the generated object tokens assign a large proportion of their attention weights to their visual grounding regions, as shown in Figure 2. This phenomenon effectively links generated object-related tokens to precise regions in the input image and is notably absent in the earlier and final layers. This insight allows us to pinpoint where the model “looks” in the input image while generating text tokens.

Equipped with this analytical tool, we found that real objects tend to focus their attention on their corresponding image anchor regions. In contrast, hallucinatory objects, lacking a true image anchor, concentrate their attention on visually related regions. For example, a nonexistent *baseball glove* could focus its attention on the *baseball* in the image. This finding highlights the critical role of visual bias in generating object hallucinations. When the image anchor of a real object is masked, the model fails to recognize the object. Inspired by this, we apply a *zero-out* operation to the top-K image tokens with the highest attention scores for hallucinated objects, replacing them with zero embeddings. Surprisingly, these hallucinated objects disappear from the new response. We further identified certain image tokens among the top-K zeroed-out tokens—termed **Hallucinatory Image Tokens (HITs)**—that play a key role in generating object hallucinations. As shown in Figure 1, the hallucinated object could be eliminated by removing even two or three HITs. In contrast, we found that real objects remain largely unaffected by the zero-out operation. This distinction shows a promising way to both detect and mitigate object hallucination.

To evaluate the generalizability of this finding, we curated a dataset of hallucination cases and applied the top-K candidate HITs zero-out operation. The results show that we can successfully reduce most OHs while preserving accurate predictions for real objects. We extended these findings and proposed a novel method, `EAZY`, which can detect and **E**liminate object h**A**llucination by **Z**eroing-out hallucinator**Y** image tokens. `EAZY` first identifies hallucinated objects in the generated text by removing object-specific top-K candidate HITs, with a detection rate up to 80%. After the hallucination detection stage, the `EAZY` will rectify the generated response by zeroing out the aggregated HIT candidates from all the detected HOs. Our experiment results show that `EAZY` surpassed the performance of state-of-the-art (SOTA) training-free methods.

Our contributions can be summarized as below:

- We show that the LVLMs extract object-related information from the visual input in the early middle to late middle layers. This relationship builds strong connections between visual regions and generated object tokens, allowing object localization via attention distribution.
- We discovered an interesting and significant phenomenon: Most HOs generated by LVLMs are directly linked to only a subset of image tokens receiving high attention scores. By zeroing out these *hallucinatory image tokens*, the associated HOs can be eliminated.
- By further validation of the HITs phenomenon, we found that zeroing out top-K candidate HITs not only removes

the majority of HOs but has minimal impact on real objects. We propose to use the HITs phenomenon for OH detection, achieving nearly 80% accuracy and precision for both real and hallucinatory objects.

- We propose a method named `EAZY` for detecting and mitigating OH by zeroing out candidate HITs. Our experimental results demonstrate that the proposed method can adapt to various LVLMs and outperforms state-of-the-art baseline approaches across most evaluation metrics.

2. Connecting Generated Object Token with Its Image Anchor via Attention

We consider LVLMs represented by LLaVA[20], with a model architecture consisting of a vision encoder, a vision-text connecting module, and a transformer-based large language model as the text decoder. An input image I is first divided into N fixed-size patches and then processed by ViT f_V and a linear projector or MLP into a sequence of image tokens. For LLaVA-1.5, a single image will be mapped as 576 image tokens for the LLM decoder.

Given an input sequence $\mathbf{x} = (x_0, x_1, \ldots, x_{M+N-1})$, the text decoder $p(x_t|\mathbf{x}_{<t}) = f_T(\mathbf{x}_{<t}|\theta)$ gives the probability distribution of next token over the vocabulary space. Here, M is the text token number; θ are the model parameters; $\mathbf{x}_{<t}$ is the input sequence that contains the tokens prior to the t-th token x_t. The complete generated token sequence $\mathbf{x}_{output}$ is obtained via $F_T(\mathbf{x}|\theta)$. We denote $\mathbf{x}_V$ as input image tokens and $\mathbf{x}_T$ as the input text tokens.

For a token x_i, its representation $h_i^l \in \mathcal{R}^{d_m}$ at the layer l of the text decoder is updated through multi-head self-attention (MHSA) and feed-forward(FF) sublayers with residual connection. Specifically, for the MHSA sublayer with H heads, each head applies self-attention as:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_h}} + M)V \qquad (1)$$

Here $Q, K, V \in \mathcal{R}^{n \times d_k}$ are the query, key, and value matrices linearly projected from the input; $M \in \mathcal{R}^{n \times n}$ is the casual mask. We utilize the average softmax output of the self-attention $a^l = \text{softmax}(\frac{QK^T}{\sqrt{d_h}} + M)$ as the attention scores at layer l. We denote the casual self-attention scores from the token x_i to x_j at layer l as a_{ij}^l.

2.1. Evaluating Object Localization Performance among Model Layers.

Understanding how LVLMs process visual input is still in its early stages. Recent studies [8, 33] have preliminarily revealed that models such as LLaVA assign a higher proportion of attention weights over the image tokens in their middle to late layers. However, these analyses fail to identify a clear dependency between the generated object-related tokens and specific image regions within the cluttered, layer-by-layer attention distributions.

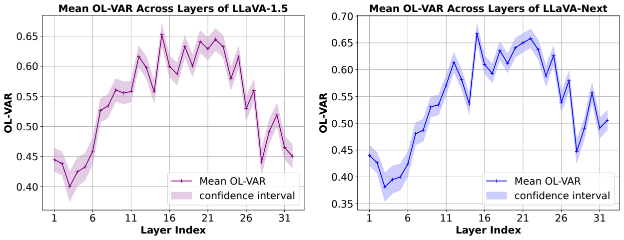


Figure 2. Average OL-VAR over 500 randomly selected MSCOCO images of LLaVA-1.5(left) and LLaVA-Next(right).

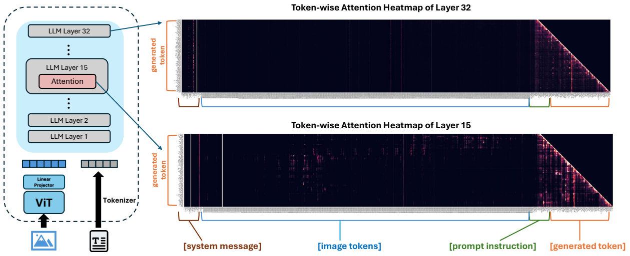


Figure 3. The token-wise attention heatmap of layer 32 and 15 from the LLaVA model. The attention distribution in layer 15 illustrates the connection between the generated object tokens and image tokens, indicating the model extracts object information from the image at the middle layers.

To explore where the model establishes connections between the generated object text tokens and their corresponding image regions, i.e., image anchors, we evaluated the performance of object localization using token-to-token attention relationships at each layer. To quantitatively assess the relationship between generated object tokens and their corresponding image regions, we introduced a new metric, the Object Localization Visual Attention Ratio (OL-VAR), to represent the ratio of attention scores over the object bounding box region to all the image tokens by generated object token x_i in layer l:

$$\text{OL-VAR}_i^l = \frac{\sum_{x_k \in \mathbf{x}_{V_O}} a_{ik}^l}{\sum_{x_j \in \mathbf{x}_V} a_{ij}^l}. \qquad (2)$$

Here, $\mathbf{x}_V$ represents the set of all the image tokens and $\mathbf{x}_{V_O}$ is the set of image tokens contained in the ground truth bounding box of the target object O. We randomly selected 500 MSCOCO images containing annotated objects and their corresponding bounding boxes and show their average OL-VAR over each layer of LLaVA-1.5 and LLaVA-Next in Figure 2. It can be seen that the model assigns more than 55% of the visual attention weights on the anchors of the object-related image from layer 10 to 25, where layer 15 obtained the highest score OL-VAR. This result offers a new perspective on understanding how LVLMs extract object information from visual inputs, revealing that the model primarily extracts and comprehends object information in the early-middle to later-middle layers. In addition, it provides an effective analytical tool for establishing connections between visual inputs and generated tokens.

We further validated this finding through the visualization of attention heatmaps. In Figure 3, we present two representative token-wise attention heatmaps of LLaVA-1.5 7b with 32 transformer layers. The top panel shows the attention distribution for Layer 32, while the bottom panel depicts the distribution for Layer 15, the layer with the highest OL-VAR on objects. For any given image input, the attention heatmaps of most layers exhibit patterns similar to those of Layer 32. Specifically, the attention scores from generated tokens to the image tend to concentrate on a few specific image tokens, forming an *attention sink* pattern [34]. In contrast, the attention heatmap of Layer 15 exhibits a markedly different distribution. The amount of attention sink pattern is significantly diminished. Instead, multiple distinct attention spots emerge, connecting the generated text tokens with various image tokens. This indicates a more distributed attention mechanism at layer 15, enabling richer interactions between text and visual elements.

3. Object Hallucination from Image Tokens.

This section presents a new perspective on the causes of hallucinations. We found that the presence of hallucinated objects in the generated text is directly correlated with the few image tokens that receive the highest attention scores. We illustrate more details about the distribution of text-to-image token attention in Appendix H.

3.1. Identify Hallucinatory Image Tokens

We begin with a case study where the model is prompted to describe an image of *“a fruit stand with bananas and kiwis"*, as shown in Figure 1. The discovery in Section 2 provides a tool for connecting generated object tokens with their visual grounding. Erasing the image anchor region of a real object prevents the model from recognizing its presence. Inspired by this, we selected the top five image tokens that received the highest attention scores for the hallucinatory objects "apple" and "orange" in the original generation and replaced them with zero embeddings (14 image tokens from multiple HO tokens). Surprisingly, "apple" and "orange" disappeared, while the real object "kiwi" emerged.

Our further analysis aims to identify the smallest set of image tokens capable of eliminating hallucinations. Ultimately, we identified three key image tokens among the top-attended tokens, located in the lower part of the ’papaya’ region. Remarkably, zeroing out just these three tokens achieved the same effect as removing all 14 image tokens with high attention scores. To further investigate how the LLM interprets these image tokens, we applied the Logit Lens [25] technique, which maps the hidden states of image tokens to the LLM’s vocabulary space. In Figure 4, we show the top-5 projected vocabulary words associated with several image tokens in the papaya region. Surprisingly, the three key image tokens are interpreted as "apple" and "orange" by Logit Lens. This aligned with our initial observation that certain image tokens are directly associated with hallucinatory objects because of inherent visual biases. We define this type of image tokens as the ***Hallucinatory Image Tokens*** (HITs)[1]. In the next section, we demonstrate that zeroing out the top-K candidate HITs (top-K most attended image tokens by HOs) can universally reduce object hallucination across different images and contexts.

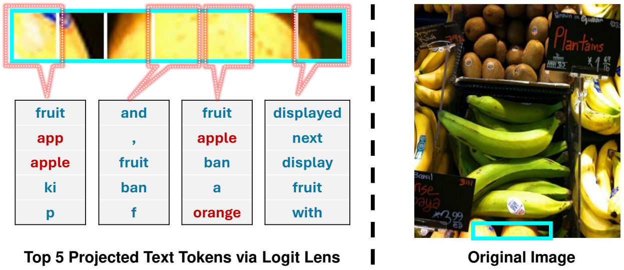


Figure 4. Logit Lens [25] interpretation of HITs. We display the top-5 projected words of image tokens related to HOs. The HITs are interpreted as the hallucinatory objects, "apple" and "orange".

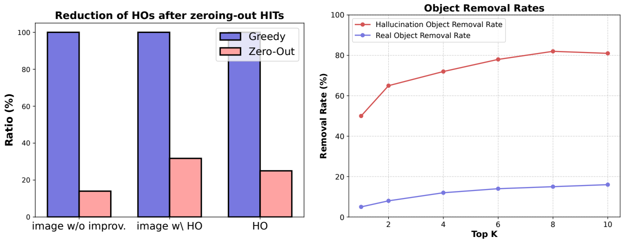


Figure 5. **Left:** Reduction ratio of OHs in Hall-COCO after zeroing out HITs. We show the ratio changes of image w/o improv.(ratio of images with the same HOs after zeroing out), image w OH (images with OH in the response), and HO (Ratio of HOs in all text responses). **Right:** The fraction change of real and hallucinatory objects removed by zeroing out the top-K candidate HITs.

3.2. Quantifying the Impact of Zeroing Out HITs

To further validate that the HITs phenomenon is broadly applicable, we curated a dataset named *Hall-COCO*, containing 200 images from MSCOCO dataset [28] that consistently induce object hallucinations identified by human annotation. We provide more details in Appendix B.

---

[1] Logit Lens alone cannot capture all HITs. Interpretability-based methods for detecting such tokens could offer improved alignment, and we plan to consider alternate tools to refine this approach in the future.

We evaluate how zero-out the candidate HITs (top attended image tokens by HO tokens) will impact the number of object hallucinations in the model's responses. As presented in Figure 5 Left, we can see that after zero-outing the top 5 candidate HITs with the same greedy decoding, 86% of images got improved description with less hallucination. The proportion of images with hallucination decreased from 100% to 31.73%. 74.17% of the object hallucination directly disappeared from the updated generation. This verified that most hallucinatory image tokens are part of the image tokens that received the most attention weights from the OH tokens. Removing these top-K candidate hallucinatory image tokens leading to the removal of OH can be widely applied to different images.

3.3. Real Objects are Robust to Zero-Out

We demonstrate that zeroing out hallucinatory image tokens significantly reduces the generation of hallucinated object tokens. This raises an interesting question: What is the effect of zeroing out high-attention patches for real objects? To explore this, we evaluate the impact of zeroing out the top-K HITs on both hallucinated and real objects. Specifically, we measure the fraction of objects removed by counting how many objects remain in the model response after removing the top-K image tokens with the highest attention scores from the corresponding text object token, relative to the total number of objects. The results, shown in Figure 5 (Right), reveal that object hallucinations are highly sensitive to zeroed HITs. For example, when $K = 8$, approximately 80% of hallucinated objects were successfully eliminated, while only a small fraction (10%) of real objects were affected. Notably, the few real objects that were removed tended to be small or positioned at the periphery of the image. We provide a failure case study in Appendix I.

4. Detecting and Mitigating Object Hallucination by HIT Zero-outs

In this section, we introduce our proposed methods, `EAZY` (Eliminate object hAllucination by Zero-out hallucinatorY image tokens), an automatic approach to detect and mitigate object hallucinations in LVLMs.

4.1. Object Hallucination Detection

Our results in Section 2 and 3 indicate that removing the top-K candidate HITs from consideration effectively eliminates HOs in the revised model output, while largely preserving the integrity of real objects. Leveraging this differentiation between genuine and HO tokens, we suggest identifying OH by discarding the top-K candidate HITs.

Given an image input I and its corresponding input token sequence $\mathbf{x}_I$, the text decoder produces G generated tokens $\mathbf{x}_{output} = \{x_{M+N}, \ldots, x_{M+N+G-1}\}$. We then utilize pre-built NLP techniques such as POS tagging and dependency parsing [2] to extract object tokens in the generated response $\mathbf{x}_{output}$, resulting in a subset $\mathbf{x}_{obj}$. For any $x_i \in \mathbf{x}_{obj}$, let a_{ij} represent the attention score of x_i on an image token x_j. The zero-out token set $\mathbf{z}_i$ is defined as

$$\mathbf{z}_i = \{x_j \in \mathbf{x}_I \mid j \in \arg\operatorname{top}_K (\{a_{ij} \mid x_j \in \mathbf{x}_I\})\}, \quad (3)$$

where $\arg\operatorname{top}_K$ selects the indices of the K largest attention scores $\{a_{ij} \mid x_j \in \mathbf{x}_I\}$. We modify the image input by applying zero-out based on $\mathbf{z}_i$,

$$\mathbf{x}_{input}^{\mathbf{z}_i} = \{x_j \leftarrow \mathbf{0}_{d_m} \mid x_j \in \mathbf{z}_i\}. \quad (4)$$

Here, $\mathbf{x}_{\text{input}}^{\mathbf{z}_i}$ indicates the input sequence with the tokens in $\mathbf{z}_i$ replaced by zeroing embedding $\mathbf{0}_{d_m} \in \mathcal{R}^{d_m}$. We check the updated generation result $\mathbf{x}_{output}^{i}$ by text decoder $\mathbf{x}_{output}^{i} = F_T(\mathbf{x}_{\text{input}}^{\mathbf{z}_i} \mid \theta)$. If $x_i \notin \mathbf{x}_{\text{output}}^{i}$, `EAZY` classifies x_i as an object hallucination, and a real object otherwise.

Algorithm 1 `EAZY`

```
Require: input tokens $\mathbf{x}_{input}$, text decoder $F_T$, attention scores $a_{ij}$, top-$K$ threshold $K$, input image $I$
Ensure: Mitigated response $\mathbf{x}_{output}$
Compute initial response: $\mathbf{x}_{output} \leftarrow F_T(\mathbf{x}_{input} \mid \theta)$
Extract object tokens: $\mathbf{x}_{obj} \leftarrow \mathbf{x}_{output}$
Identify top-$K$ attended image tokens for all objects:
    $\mathcal{Z}_I^{estimate} = \bigcup_{x_i \in \mathbf{x}_{obj}} \arg\operatorname{top}_K(\{a_{ij} \mid x_j \in \mathbf{x}_{obj}\})$
Zero-out all hallucinated tokens: $\mathbf{x}_{input}^{\mathcal{Z}_I^{estimate}} \leftarrow \mathbf{x}_I$ with $\mathcal{Z}_I^{estimate}$ replaced by $\mathbf{0}_{d_m}$
Generate updated response: $\mathbf{x}_{output}^{HIT} = F_T(\mathbf{x}_{input}^{\mathcal{Z}_I} \mid \theta)$
Identify hallucinated objects: $\mathbf{x}_{obj}^{HO} = \{x_i \in \mathbf{x}_{obj} \mid x_i \notin \mathbf{x}_{output}^{HIT}\}$
Final zero-out of hallucinated tokens:
    $\mathcal{Z}_I^{final} = \bigcup_{x_i \in \mathbf{x}_{halluc}} \arg\operatorname{top}_K(\{a_{ij} \mid x_j \in \mathbf{x}_I\})$
Generate final response: $\mathbf{x}_{output}^{\text{EAZY}} = F_T(\mathbf{x}_{input}^{\mathcal{Z}_I^{final}} \mid \theta)$
    return $\mathbf{x}_{output}^{\text{EAZY}}$
```

4.2. Mitigating Object Hallucination

As shown in Figure 6, the mitigation application of `EAZY` starts from an initial inference without intervention. Following a similar process as the detection application, the object tokens x_{obj} are extracted from the initial response. We then group the top-K candidate HITs $\mathcal{Z}_I^{estimate}$ for all objects $x_i \in \mathbf{x}_{obj}$ and apply zero-out to the image tokens accordingly. The model then performs inference for HITs estimation after zeroing out the corresponding image tokens.

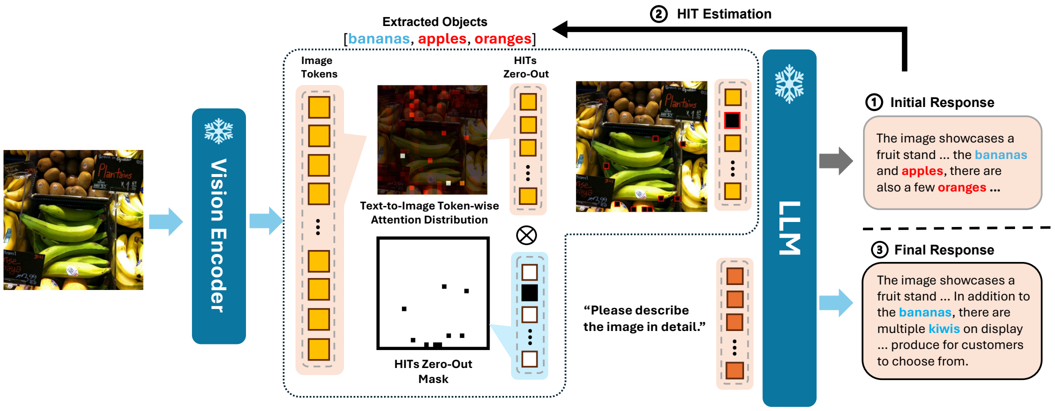


Figure 6. Overview of the proposed `EAZY` method. The process starts with an image input encoded into image tokens. The LVLM generates an initial response with hallucinated objects (apples, oranges). `EAZY` estimates HITs via text-to-image token-wise attention distribution. With HITs zeroed out, the final response has the hallucinations disappeared, revealing correct objects (kiwis).

`EAZY` identifies objects that disappear in the new response as hallucinatory object tokens $\mathbf{x}_{obj}^{HO}$.

We collect all the top-K candidate HIT sets of the HO tokens detected from the previous HIT estimate stage. We take the union set of all the candidate HITs sets to obtain the final zero-out token list $\mathcal{Z}_I^{final}$, which is defined as

$$\mathcal{Z}_I^{final} = \bigcup_{x_i \in \mathbf{x}_{obj}} \mathbf{z}_i \quad \text{where} \quad x_i \in \mathbf{x}_{obj}^{HO}. \tag{5}$$

The model then utilizes the final zero-out token list $\mathcal{Z}_I^{final}$ to replace all the candidate hallucinatory image tokens by zero embedding for the final generation:

$$\mathbf{x}_{output}^{\texttt{EAZY}} = F_T(\mathbf{x}_{input}^{\mathcal{Z}_I^{final}} \mid \theta). \tag{6}$$

5. Experiment

In this section, we conduct experiments on multiple datasets and metrics to validate the superior performance of `EAZY` in detecting and mitigating hallucinated objects (HOs).

5.1. Experiment Setting

5.1.1. OH Mitigation Setting

Model. We choose the LLaVA-1.5 [18], Shikra [3] at 7B scale and LLaVA-Next [19] at 8B scale for evaluation of OH mitigation. We set the maximum number of generated tokens as 512 in our setting. Please refer to Appendix C.1 for more model details.

Baseline. We compare our method `EAZY` with six *training-free* methods including three LLM decoding strategies, greedy and beam search and Dola [5], as well as three state-of-the-art hallucination mitigation methods for LVLM, VCD [14], OPERA [10] and SID [11]. We adopt greedy decoding for Dola, SID and EAZY. We set the beam number as 5 and apply sampling for VCD and OPERA. We provide more implementation details in Appendix C.3.

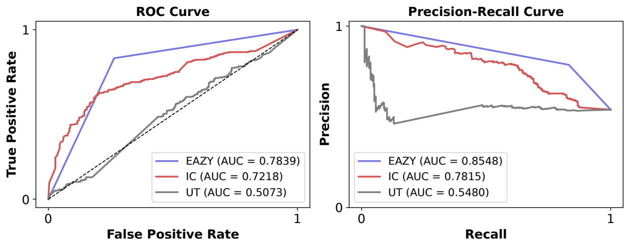


Figure 7. **Object Hallucination Detection Curves on Hall-COCO.** We present the Precision-Recall and ROC curves of the proposed OH detection method and baselines.

5.1.2. OH Detection Setting

Baselines. We chose two state-of-the-art *training-free* OH detection methods as baselines to compare our method on Hall-COCO. Uncertainty (UT) [36] found that HOs often exhibit higher uncertainty scores during generation. Internal Confidence (IC) [12] found the maximum logit lens probability can be utilized for OH detection.

Model and Benchmark. We evaluate different OH detection methods on the LLaVA-1.5 7B model with greedy decoding. The maximum length of the generation is 512 tokens. We use Hall-COCO as the benchmark for the open-end image description task. Each image has human-annotated real and hallucinatory objects.

Metric. We follow [12] to treat the OH detection as a binary classification task, where real objects are the positive class and HOs as the negative. We use accuracy (Acc), precision (PR), recall, and F1 score as the evaluation metrics.

Method	**LLaVA-1.5**		**Shikra**		**LLaVA-NEXT**	
	CHAIR$_S$ ↓	CHAIR$_I$ ↓	CHAIR$_S$ ↓	CHAIR$_I$ ↓	CHAIR$_S$ ↓	CHAIR$_I$ ↓
Greedy	49.6	14.4	55.8	15.4	32.8	9.1
Beam Search	46.3	12.9	50.4	13.3	33.0	9.2
Dola	47.1	13.8	55.4	15.7	31.3	9.0
VCD	49.2	14.8	56.4	15.5	32.8	8.9
OPERA	45.4	12.7	46.2	13.1	34.0	8.7
SID	44.2	12.2	44.8	12.8	32.8	8.3
EAZY	**38.8**	**11.4**	**26.6**	**8.9**	**26.8**	**8.3**

Table 1. **Evaluation Results with CHAIR.** The lower, the better.

Method	**LLaVA-1.5**		**Shikra**		**LLaVA-NEXT**	
	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑
Greedy	81.38	82.20	80.87	81.05	83.78	82.24
Beam Search	84.66	84.60	80.48	81.81	83.77	81.69
Dola	84.06	84.62	80.71	81.19	84.53	84.77
VCD	84.66	84.52	79.87	81.06	83.37	83.95
OPERA	84.88	85.21	79.99	81.48	84.37	84.21
SID	84.82	85.50	80.02	81.23	84.61	85.32
EAZY	**84.97**	**85.78**	**80.96**	**82.38**	**84.91**	**85.40**

Table 2. **Evaluation Result on POPE.** We take the average accuracy and F1 score of random, popular, and adversarial models.

Metric	**UT**	**IC**	**EAZY**
Acc	50.57	62.56	**78.77**
PR(RO)	53.60	61.93	**78.41**
PR(OH)	42.77	64.16	**79.25**
Recall	70.62	81.60	**83.38**
F1	60.95	70.42	**80.82**

Table 3. **OH Detection Results on Hall-COCO.** PR(RO) represents the precision of real objects (positive instances), while PR(OH) represents the precision of object hallucination (negative instances). **UT** [36] is the uncertainty detection method. **IC** [12] is the internal confidence method.

5.2. Evaluation

5.2.1. Evaluation Result on OH Mitigation

We adopt CHAIR [28] and POPE [16] as quantitative metrics to compare the hallucination mitigation performance between the proposed EAZY and baseline methods. We provide detailed descriptions of both metrics in Appendix C.2. For the **CHAIR** result in Table 1, we follow [6, 8, 10, 11] to evaluate different LVLMs with 500 randomly selected images from MSCOCO [17] with the fixed prompt: *"Please describe this image in detail."*. The proposed method EAZY surpasses the baselines by a large margin for three different LVLMs, validating the effectiveness of EAZY in open-end generation tasks. We show the average evaluation results on POPE in Table 2, which comprises three datasets. EAZY attains the best average performance in random, popular, and adversarial sampling settings, outperforming all other methods. It can be observed that, under the POPE metric, the advantage of EAZY is not as pronounced as in the CHAIR results. We attribute this to the binary Yes-or-No question format of POPE, which may lead to generated content that does not include object tokens. Furthermore, the prompt *"Is there an object in the image?"* already contains the object name, potentially impacting models' attention distribution.

5.2.2. Evaluation Result on OH Detection

We present the quantitative results of OH detection on LLaVA-1.5 in Table 3 and Figure 7. We find that EAZY improves the accuracy 16.21%, 16.48% in the precision of real objects, and 15.09% in the precision of HOs, compared to the best baseline. In addition, EAZY improves the ROC AUC by 8.60% and the PR AUC by 9.37% over the baseline. This indicates the superior performance of EAZY on OH detection with a good balance of precision on both real and hallucinatory objects. We provide additional results for LLaVA-Next and Shikra models in Appendix C.5.

5.3. Evaluation on MLLM benchmarks.

We evaluate EAZY on on two popular MLLM benchmark, MME [7] and MMBench [23]. Based on the result in Table 4, we can see EAZY maintains the MLLM's general capability on both benchmarks.

Table 4. MME & MMBench result of LLaVA-1.5-7b

Method	Greedy	Dola	VCD	OPERA	SID	`EAZY`
MME	1510.7	1480.1	1488.7	1515.4	1520.2	1496.2
MMBench	64.3	63.8	63.9	64.4	65.1	64.1

5.4. Ablation Study

Impact of K value selection. The choice of the value K is critical to the mitigation performance of `EAZY`. In Figure 8, we show the impact of different values K for the LLaVA-1.5 and Shikra models on the CHAIR metric. We use the Greedy decoding performance as the reference line. It can be seen that LLaVA-1.5 achieves the best performance when $K = 5$ and Shikra is on $K = 3$.

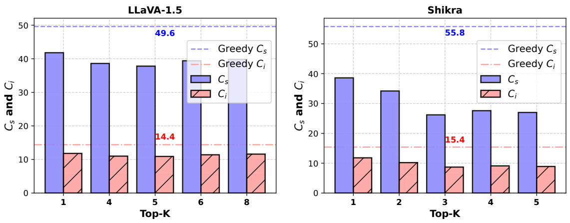


Figure 8. CHAIR evaluation results as a function of K on LLaVA-1.5 (left) and Shikra (right).

Position	CHAIR$_S$ ↓	CHAIR$_I$ ↓
Input Token	38.8	11.4
ViT Output	40.4	11.8
Image Patches	39.0	11.4

Table 5. Evaluation Result of Different Zero-Out Positions on CHAIR.

Impact of Zero-Out Position in the Pipeline. In `EAZY`, we apply the zero-out of image tokens in the LLM input space, i.e., replace the target image tokens with zero embeddings. To transform the original image into image tokens, the LVLMs first utilize the ViT to process the image from image patches into patch embeddings, then use the linear projector or MLP to map the patch embeddings into image tokens. We apply the zero-out operation on three different positions in the image processing pipeline, including the input token space, the output space of ViT, and the original image patches. As shown in Table 5, applying the zero-out operation at different positions on the same set of image tokens (or patch embeddings) effectively mitigates object hallucination in the model response. This suggests that the visual biases introduced by hallucinatory image tokens originate directly from the image itself and are not strongly associated with modules such as ViT or Linear Projector.

Methods	**Image w/o OH**	**Removed OH**
Zero-Out	**68.27%**	**74.17%**
Average	55.17%	66.67%
Random	58.56%	63.33%
Neighbor	55.17%	61.25%

Table 6. Evaluation Result of Top-K Image Token Processing Strategies on Hall-COCO.

Impact of Image Token Processing Strategies. We explored how the different processing methods for the top-K image tokens impact the removal of object hallucination (OH). In Table 6, we show the performance of four different processing strategies on the collected *Hall-COCO* dataset when $K = 5$. The zero-out method, which is used in `EAZY`, replaces the embeddings of the top-K candidate HITs with zero embeddings. The average method replaces the top-K most attended image token embeddings with the average embedding of all the image tokens. The random method replaces the token embeddings of top-K candidate HITs with a randomly selected image token embedding. The neighbor method replaces the top-K candidate HITs embeddings with one of their neighbor image tokens adjunct to the target image token. Overall, our zero-out performs better in the percentage of images with fewer object hallucinations, images without object hallucinations, and removed object hallucinations compared to alternative strategies.

6. Conclusion

In this paper, we identify a new cause of object hallucination from the perspective of visual bias—Hallucinatory Image Tokens. A small number of image tokens with high attention scores are directly correlated with the presence of hallucinated objects in the generated text. Replacing these HITs with zero embeddings effectively removes hallucinated objects while minimally affecting real objects. Based on this finding, we propose `EAZY`, a training-free method for automatically detecting and mitigating object hallucinations. Through extensive experiments, we demonstrate the superior performance of `EAZY` in both hallucination detection and mitigation. We provide more discussion and visualization results in Appendix J and K.

References

[1] 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. 11

[2] Steven Bird. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72, 2006. 5

[3] 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. 6, 11

[4] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023. 1, 11

[5] 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 *The Twelfth International Conference on Learning Representations*, 2024. 6

[6] Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14303–14312, 2024. 7, 11

[7] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024. 7

[8] Xuan Gong, Tianshi Ming, Xinpeng Wang, and Zhihua Wei. Damro: Dive into the attention mechanism of lvlm to reduce object hallucination. *arXiv preprint arXiv:2410.04514*, 2024. 2, 3, 7, 11

[9] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024. 1

[10] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multimodal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427, 2024. 1, 6, 7, 11

[11] Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-introspective decoding: Alleviating hallucinations for large vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025. 6, 7

[12] Nick Jiang, Anish Kachinthaya, Suzie Petryk, and Yossi Gandelsman. Interpreting and editing vision-language representations to mitigate hallucinations. *arXiv preprint arXiv:2410.02762*, 2024. 6, 7, 13

[13] Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. Volcano: mitigating multimodal hallucination through self-feedback guided revision. *arXiv preprint arXiv:2311.07362*, 2023. 1, 11

[14] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882, 2024. 1, 6, 11, 14

[15] Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. *arXiv preprint arXiv:2210.15097*, 2022. 1, 11

[16] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 1, 7, 11

[17] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 7, 11, 12

[18] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 1, 6

[19] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 6, 11

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1, 3, 11

[21] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024. 1

[22] Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, pages 125–140. Springer, 2025. 11

[23] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2024. 7

[24] Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. *arXiv preprint arXiv:2410.07149*, 2024. 2

[25] nostalgebraist. Interpreting gpt: The logit lens, 2020. Accessed: 2024-12-21. 4

[26] OpenAI and Achiam et al. Gpt-4 technical report, 2024. 11

[27] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 11

[28] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018. 1, 4, 7, 11

[29] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023. 1

[30] 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. *arXiv preprint arXiv:2311.07397*, 2023. 11

[31] 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. 1

[32] Jinfeng Wei and Xiaofeng Zhang. Dopra: Decoding over-accumulation penalization and re-allocation in specific weighting layer. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7065–7074, 2024. 1, 11

[33] Sangmin Woo, Donguk Kim, Jaehyuk Jang, Yubin Choi, and Changick Kim. Don't miss the forest for the trees: Attentional vision calibration for large vision language models. *arXiv preprint arXiv:2405.17820*, 2024. 2, 3, 11

[34] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023. 4

[35] Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12944–12953, 2024. 11

[36] 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. *arXiv preprint arXiv:2310.00754*, 2023. 1, 6, 7, 11, 13