A. Related Work

Cause of Object Hallucination. Hallucination [1] affects the trustworthiness and reliability of LVLM for broad application. Specifically, object hallucination [28], which refers to describing HOs in the input images, is more harmful than other types of inaccurate description. Recent studies have proposed explaining the causes of hallucination from training data biases [30, 35, 36] and language biases [6, 10, 13, 14]. In [6, 22], the authors found that LVLMs exhibit a preference for text tokens, causing the model to pay more attention on neighboring text tokens when generating new tokens.

Mitigation of Object Hallucination. Due to the complex structure and the large number of parameters in LVLMs, addressing hallucinations through retraining or fine-tuning is often prohibitively expensive and computationally intensive. As a result, many studies focus on mitigation during inference. A series of decoding strategy-based works [6, 14, 32, 33] leverage contrastive decoding technique [15] to compare the model's original logit probability distribution with the logit output generated from processed images, thereby suppressing prior knowledge biases and statistical biases introduced by the training data. Meanwhile, methods such as OPERA [10] and DAMRO [8] have observed that sequence supervision-induced anchor tokens significantly increase the likelihood of hallucinations in subsequent tokens. Compared to these works, we discovered that visual biases are the cause of most object hallucinations and developed a simple yet effective method by zeroing out the hallucinatory image tokens.

B. Hall-COCO: Benchmarking for Object Hallucination Mitigation Evaluation

To validate that the phenomenon of HITs can be broadly applied to other images, we constructed a dataset, named *Hall-COCO*, from MSCOCO [17] where each image reliably induces object hallucination. The dataset creation involved several steps. First, we used GPT-4o [26] to detect hallucinations in the output of LLaVA. Next, human annotators identified specific object hallucinations in the model's responses. For each hallucinated object, we then identified the top 5 image tokens with the highest attention scores. In cases where multiple hallucinated objects were present in the output, we applied a union operation to compile the set of relevant tokens. Through this process, we curated a dataset containing 200 images that consistently induce object hallucinations.

Image	Real Objects	OH
200	333	284

Table 7. The number of images, real objects and object hallucinations (OH) in Hall-COCO.

C. Extended Experiment

C.1. Model Details.

LLaVA-1.5. The LLaVA-1.5 Model [20] leverages the linear projector layer to align the vision and text modalities, with 576 image tokens. It adopted the pre-trained vision transformer from CLIP[27] and the pre-trained language model as Vicuna[4].

Shikra. The Shikra Model [3] introduces referential dialogue capabilities in multimodal large language models (MLLMs) by handling spatial coordinate inputs and outputs in natural language. It utilizes a vision encoder, an alignment layer, and a Vicuna-based language model without requiring extra vocabularies, position encoders, pre-/post-detection modules, or external plug-ins. The model enables interaction with images through natural pointing and location-aware responses, supporting tasks such as Referring Expression Comprehension (REC), PointQA, Image Captioning, and Visual Question Answering (VQA) with promising performance.

LLaVA-Next. The LLaVA-Next Model [19] enhances multimodal capabilities by increasing the input image resolution up to 672 × 672 pixels, supporting three aspect ratios. It utilizes an improved visual instruction tuning data mixture to bolster visual reasoning and OCR capabilities. The model employs a pre-trained vision transformer from CLIP and integrates with advanced language models like Vicuna and Mistral.

C.2. Evaluation Metric of OH Mitigation

CHAIR The Caption Hallucination Assessment with Image Relevance (CHAIR) [28] is an evaluation tool designed to assess object hallucination issues in the image captioning task. CHAIR quantifies the degree of object hallucination from sentence-level,CHAIR$_S$, and image-level, CHAIR$_I$, which are calculated by the ratio of hallucinaty objects and the ground-truth label objects. Specifically, CHAIR is calculated as:

$$\text{CHAIR}_S = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all mentioned objects}\}|}, \tag{7}$$

$$\text{CHAIR}_I = \frac{|\{\text{captions w/ hallucinated objects}\}|}{|\{\text{all captions}\}|} \tag{8}$$

POPE Polling-based Object Probing Evaluation (POPE) [16] is a metric designed to evaluate object hallucination in LVLMs. During evaluation, POPE prompts

LVLM with straightforward Yes-or-No questions about the presence of specific objects in an image (e.g., "`Is there an apple in the image?`") POPE offers three evaluation settings: *Random, Popular, and Adversarial*. In Random Sampling, objects that do not exist in the image are selected randomly. In Popular Sampling, the top-k most frequent objects in the entire image dataset that are absent from the current image are selected, where $k = [n/2]$. Lastly, Adversarial Sampling ranks objects by their co-occurrence frequency with ground-truth objects and selects the top-k frequent ones that are not in the image. In our experiment, we show the average accuracy and F1 score result over three different sampling modes on the MSCOCO dataset [17].

C.3. Implementation Detail

For LLaVA-1.5, we use the attention score from layer 15 (index starting from 1) to choose the top-5 most attended image tokens as the candidate HITs. For Shikre, we use the attention score from layer 18 (index starting from 1) to choose the top-3 most attended image tokens as the candidate HITs. For LLaVA-Next, we use the attention score from layer 15 (index starting from 1) to choose the top-5 most attended image tokens as the candidate HITs.

C.4. Better OH Mitigation via Iterative Hallucinatory Objects Detection

In EAZY, for each text object tokens, we take their top-K most attended image tokens and take the union of all such sets as the zero-out list for HO detection inference. However, we realize that if we detect the hallucinatory objects iteratively, EAZY can achieve better hallucination mitigation performance. Specifically, assuming we extracted n text object tokens from the initial generation, we may take each of such top-K image tokens to zero out and observe if the corresponding object disappears from the new response. In such a case, candidate HITs from different text object tokens do not interfere with each other, enabling more effective detection and mitigation of hallucinations. We denote this iterative detection variant as EAZY (iterative) and represent its performance comparison with EAZY in Table 8. It can be observed that EAZY (iterative) generally improved the performance on CHAIR with MSCOCO dataset.

C.5. Additional Evaluation Result of OH Detection

We provide the OH detection evaluation result of LLaVA-Next and Shikra in this section.

As shown in Tables 9 and 10, applying the EAZY method to LLaVA-Next and Shikra for hallucinated object (HO) detection demonstrates that EAZY continues to achieve superior performance across the majority of key metrics. In particular, EAZY exhibits significant advantages in overall accuracy and the precision of hallucinated object detection. Compared to the best-performing baseline, EAZY improves accuracy by 7.65% and 12.98% on LLaVA-Next and Shikra, respectively, while achieving gains of 46.98% and 23.56% in PR(OH). This highlights its effectiveness in correctly identifying hallucinated objects.

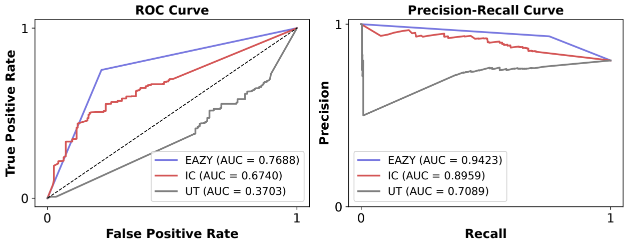


Figure 9. **Object Hallucination Detection Curves of LLaVA-Next on Hall-COCO.** We present the Precision-Recall and ROC curves of the proposed OH detection method and baselines.

On the other hand, we observe that EAZY's performance in detecting real objects is less pronounced for both models. We attribute this to differences in how LLaVA-Next and Shikra process image-to-image token transformations compared to LLaVA-1.5. Specifically, Shikra maps 576 image patch embeddings into 256 image tokens through an MLP, which causes visual information to intertwine and interfere with one another. Consequently, zeroing out specific image tokens may introduce inaccuracies. For LLaVA-Next, the model employs the AnyRes technique to split and resize the original image, requiring the removal of more image tokens than LLaVA-1.5. This likely impacts the detection of real objects by reducing the amount of preserved visual information.

Nevertheless, EAZY consistently demonstrates strong performance in object hallucination (OH) detection across both LLaVA-Next and Shikra, achieving the highest F1 scores of 80.12% and 76.61%, respectively. This indicates its superior ability to balance precision and recall, making it the most effective method overall. Particularly, EAZY achieves the highest recall on both models (92.67% on LLaVA-Next and 84.92% on Shikra), significantly surpassing the baseline methods. This suggests that EAZY effectively minimizes false negatives, ensuring that a larger proportion of hallucinated objects are correctly identified.

Furthermore, EAZY achieves the highest AUC scores in both ROC and Precision-Recall (PR) curve evaluations as shown in Figure 9 and 10. On LLaVA-Next, EAZY attains an AUC of 0.7688 for ROC and 0.9423 for PR, while on Shikra, it achieves 0.6903 for ROC and 0.8863 for PR. These results highlight EAZY 's ability to maintain a favorable trade-off between precision and recall, reinforcing its robustness across different vision-language models.

Method	LLaVA-1.5		Shikra		LLaVA-NEXT	
	CHAIR$_S$ ↓	CHAIR$_I$ ↓	CHAIR$_S$ ↓	CHAIR$_I$ ↓	CHAIR$_S$ ↓	CHAIR$_I$ ↓
EAZY	38.8	11.4	26.6	8.9	26.8	**8.3**
EAZY (iterative)	**37.8**	**10.9**	**26.2**	**8.7**	**25.6**	8.5

Table 8. **Evaluation Results with CHAIR between EAZY and EAZY (iterative).**

Metric	UT	IC	EAZY
Acc	49.05	64.29	**71.94**
PR(RO)	**87.75**	87.39	70.57
PR(OH)	24.68	30.51	**77.49**
Recall	42.32	64.77	**92.67**
F1	57.10	74.40	**80.12**

Table 9. **OH Detection Results using LLaVA-Next 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.

Metric	UT	IC	EAZY
Acc	49.59	56.03	**69.01**
PR(RO)	77.33	**81.87**	69.77
PR(OH)	43.41	34.79	**66.97**
Recall	55.60	50.80	**84.92**
F1	30.45	62.69	**76.61**

Table 10. **OH Detection Results using Shikra 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.

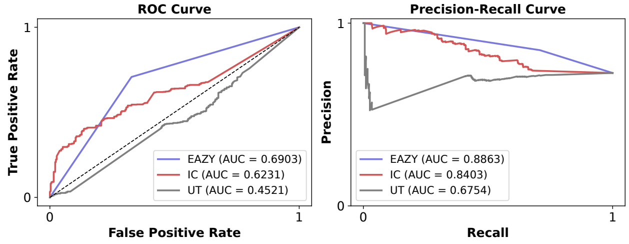


Figure 10. **Object Hallucination Detection Curves of Shikra on Hall-COCO.** We present the Precision-Recall and ROC curves of the proposed OH detection method and baselines.

D. Efficiency Analysis

We evaluate the inference time (in seconds) on the LLaVA-1.5 7B under the CHAIR evaluation setting, which is close to the real-world application scenario. As shown in Figure 11, we present the average inference time over 500 image with the prompt "Please describe the image in detail.". We evaluate the different methods on NVIDIA A6000 GPUs. It can be observed the proposed EAZY method has 50% lower inference time than OPERA, while it brings significant performance improvement on object hallucination mitigation.

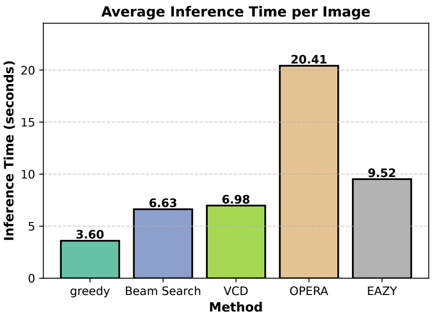


Figure 11. The average inference time of different methods on CHAIR evaluation.

E. Statistics on Image Token Zeroing-out

We measured the average number of image tokens zeroed out when applying EAZY across different LVLMs, as well as the ratio of this number to the total image tokens. As shown in Table 11, LLaVA-1.5 zeroes out an average of 8.98 image tokens out of 576, resulting in a ratio of 1.56%. Shikra removes 5.23 image tokens on average from a total of 256, yielding a slightly higher ratio of 2.04%. LLaVA-Next, which processes significantly more image tokens (2880 in total), removes an average of 40.95 tokens, with a ratio of 1.42%. These results demonstrate that EAZY requires modifying only a small fraction of image tokens to effectively mitigate hallucinations, highlighting its efficiency and minimal impact on the overall visual representation.

F. Additional Ablation Study

F.0.1. Impact of Zero-out Layer Choice

We evaluate the impact of different zero-out layer choices for hallucination mitigation. As shown in Table 12, we compare the CHAIR scores when zero-out is applied on layer

Model	LLaVA-1.5	Shikra	LLaVA-Next
# Average Zeroed-out	8.98	5.23	40.95
# Total Image Token	576	256	2880
Ratio(%)	1.56	2.04	1.42

Table 11. The statistic result of the average zeroed-out image token numbers by EAZY, the total number of image tokens and the ratio between them for each model.

Table 12. EAZY on different layers of LLaVA-1.5-7b

Layer	$\text{CHAIR}_S \downarrow$	$\text{CHAIR}_I \downarrow$
10	39.8	**11.3**
15	**38.8**	11.4
25	41.7	11.9

10, 15, and 25, respectively. It can be seen that layer 15 obtained the best performance, which aligned with our analysis in Section 2.

G. Visual Bias is the Main Cause of Object Hallucination.

The causes of Object Hallucination (OH) are complex and may include factors such as training data, model architecture, cross-modal alignment, training paradigms, and inference methods. We focus on factors that may lead to OH during the inference process, further exploring the impact of visual bias on OH and its proportion among different causes.

To simplify the problem, we categorize the factors causing OH during inference into three main categories:

- **Visual Bias**
  - Image Ambiguity: Low-resolution, misleading attributes (color, texture, shape), shadow, etc.
  - Modality Alignment: The gap between the mapped visual features and text features
- **Language Prior**:
  - Statistical learning bias of the language model leading to hallucination, including prior knowledge, grammar, co-occurrence, etc.
- **Others:**
  - Training Data Distribution Bias, Distribution imbalance hallucination, etc.

For a hallucinatory object, its emergence may result from multiple overlapping causes. If a specific factor is altered, causing the hallucinated object (HO) to disappear or change, we consider that factor to be the dominant cause of the HO.

Following with [14], we gradually add Gaussian noise on the image input to erase and weaken visual information. In such a case, we reduce the impact of visual bias while amplifying the influence of language priors and other factors. To evaluate the impact of visual bias in image inputs on object hallucination generation, we conducted experiments using LLaVA-1.5 on the Hall-COCO dataset. Specifically, we analyzed how the generation of hallucinated objects (HOs) changes when different levels of noise are added to the images. Specifically, we define New Hallucination as a newly generated hallucinatory object that appears after adding noise to the image. We define Stable Hallucination as a previously known hallucinatory object that persists even after applying noise. This suggests that Stable Hallucinations rely less on visual input and are instead more influenced by language priors.

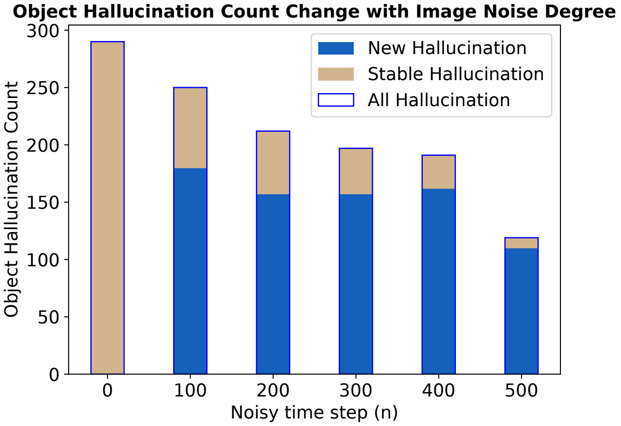


Figure 12. The amount change of different object hallucination along with the noise steps.

As shown in Figure 12, we observe that as the noise time step increases, the image gradually becomes indistinguishable from pure Gaussian noise. First, we note that the number of hallucinatory objects gradually decreases. This is primarily due to the reduction in visual information, which leads the model to generate shorter text responses.

On the other hand, we find that the proportion of New Hallucinations significantly exceeds that of Stable Hallucinations, and the proportion of Stable Hallucinations continues to decline. After simply adding noise for 100 steps, the vast majority of hallucinatory objects are New Hallucinations, indicating that visual bias is a key factor driving the variation and generation of object hallucinations.

To further estimate the relative impact of visual bias and language prior on object hallucination (OH), we fix the noise time step at 300. This setting largely removes visual information while avoiding excessive noise that would render the model's output meaningless. We track the intersection of hallucinatory objects (HOs) that persist after applying EAZY on the original image and those that remain after adding 300-step Gaussian noise. Our findings indicate that 20% of all HOs are Stable Hallucinations, among which 43% cannot be eliminated by EAZY. As shown in

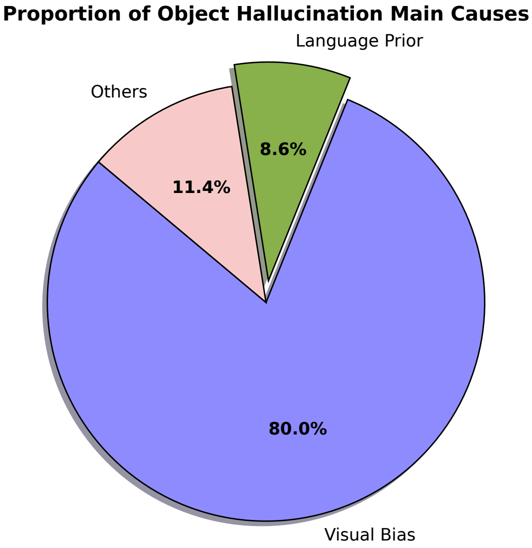


Figure 13. The proportion of different OH causes. Visual bias caused 80% of object hallucination changes.

Figure 13, this suggests that visual bias accounts for 80% of OH variations, while 8.6% of OHs are primarily influenced by language prior. The remaining 11.4% can be attributed to other factors.

H. Hallucinatory objects are looking at image regions visually close to them.

In this section, we compare the attention score distribution of real objects and hallucinatory objects over image tokens. We visualize the text-to-image token-wise attention heatmap using the attention from the 15th layer of the LLaVA-1.5 model. As shown in Figure 14, 16, 17 and 15, the attention distribution for real objects is typically concentrated on their corresponding image anchors. In contrast, hallucinatory objects follow two distinct attention patterns: One, a dispersed attention distribution across different image regions; and two, a highly concentrated focus on specific image regions that induces visual bias. Despite these differences, both patterns share a common characteristic - attended regions often have some visual resemblance to the hallucinated object.

These regions often exhibit visual similarities to the hallucinated objects. For example, in Figure 14, the attention for *apple* is distributed over the *papaya* and *kiwi* regions—*papaya* shares a similar texture and color, while *kiwi* has a comparable shape. In Figure 15, the hallucinated object *television* primarily attends to the *microwave*, likely due to the visual resemblance of its glass panel and white frame to a television. In Figure 16, the hallucinated *baseball glove* directs all its attention to the baseball itself. Similarly, in Figure 17, the hallucinated *laptop* focuses on a black square notebook on the desk, which has a shape similar to a laptop.

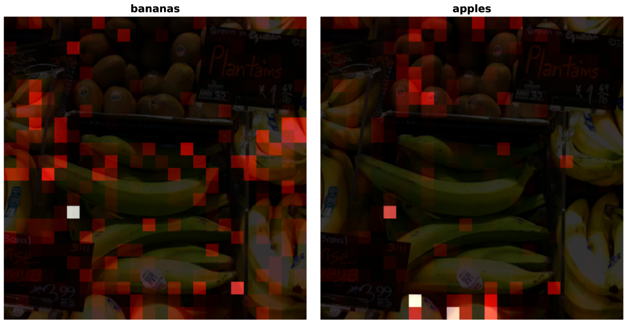


Figure 14. **Left:** The attention heatmap of the real object“*bananas*"; **Right:** The attention heatmap of the hallucinatory object“*apples*".

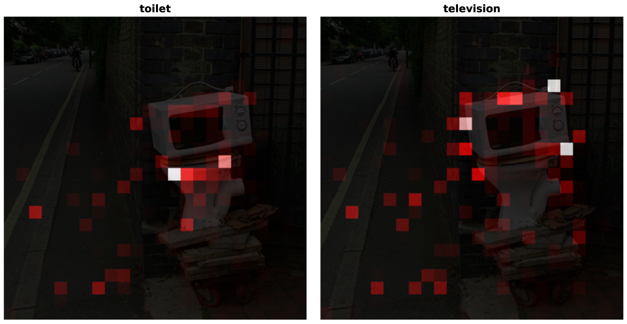


Figure 15. **Left:** The attention heatmap of the real object“*toilet*"; **Right:** The attention heatmap of the hallucinatory object“*television*".

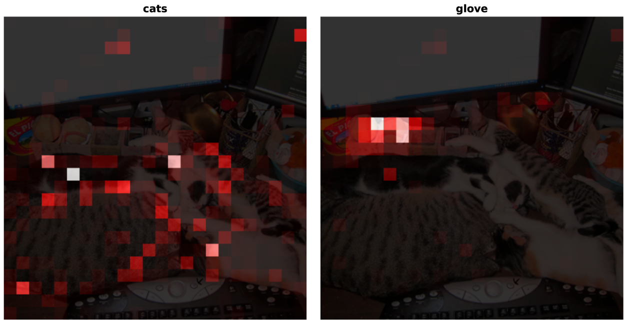


Figure 16. **Left:** The attention heatmap of the real object“*cats*"; **Right:** The attention heatmap of the hallucinatory object“*glove*".

I. Failure Case Analysis.

We identify two typical cases of real objects being removed after applying EAZY. As shown in Figure 18, the small-size

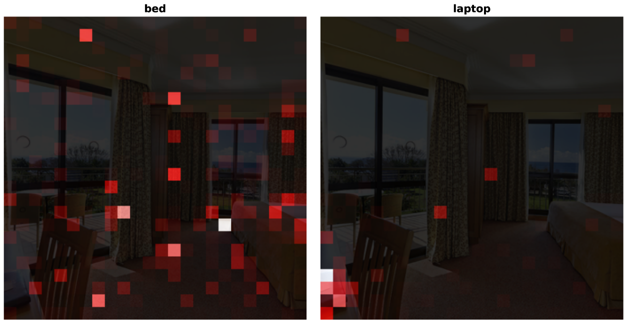


Figure 17. **Left:** The attention heatmap of the real object“*bed*"; **Right:** The attention heatmap of the hallucinatory object“*laptop*".

objects have the chance to be entirely removed from the image input, causing the model unable to recognize the objects. When the objects are located at the edge of the image with only part of them revealed, additionally, longer words are tokenized into multiple tokens, it may lead to the excessive removal of image tokens. Such an example can be seen in Figure 19.

J. Discussion and Future Work

J.1. Summary of Key Findings

Our work provides a new perspective on understanding the causes and mechanisms of object hallucination in LVLMs—visual bias. We first demonstrate that LVLMs identify and extract object-related information from image tokens in their early-middle to late-middle layers. This finding leads us to an important and intriguing phenomenon—**Hallucinatory Image Tokens** (HITs). By removing only a small subset of these tokens, we can significantly mitigate object hallucination in LVLM-generated outputs.

Building on this discovery, we propose EAZY, a training-free approach that effectively detects and mitigates OH. By replacing the top-K candidate HITs with zero embeddings, EAZY successfully suppresses hallucinated objects while preserving real object information. Our experiments demonstrated the superior performance of EAZY on HO detection and mitigation.

J.2. Limitations and Future Work

Our work leads to two promising questions. The first is achieving a deeper understanding of the HITs phenomenon. We found that existing interpretability tools, such as the Logit Lens, are insufficient for fully explaining and identifying all HITs. In future work, we aim to gain a more comprehensive understanding of HITs and leverage advanced interpretability tools to better distinguish them from their top-K candidates, thereby significantly improving the efficiency of EAZY in zeroing out image tokens.

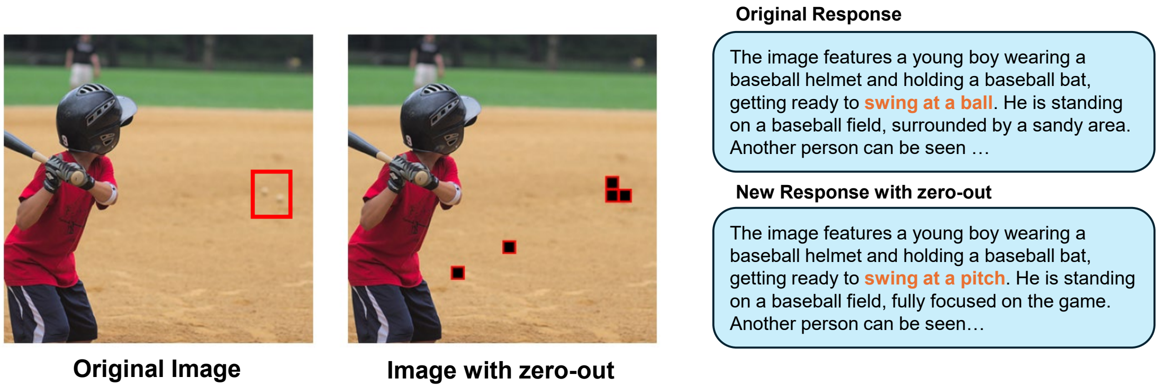
The second lies in EAZY 's handling of real objects. As discussed in Appendix I, EAZY may occasionally misclassify real objects—particularly those that are small in size, tokenized into multiple tokens, or located at the image boundaries—and erroneously remove them from the generated response. To address this, we plan to explore the following directions:

1. Performing multiple inferences to assess the robustness of different object tokens.
2. Incorporating additional features, such as uncertainty estimation, to improve the detection accuracy.
3. Integrating external vision models to refine object identification and mitigate errors.

We hope that the findings of this work will contribute to a deeper understanding of object hallucination in LVLMs, inspire future research, and enable broader applications of the HITs phenomenon in vision-language models.

K. Case Study

Please refer to the next page for more visualization cases.



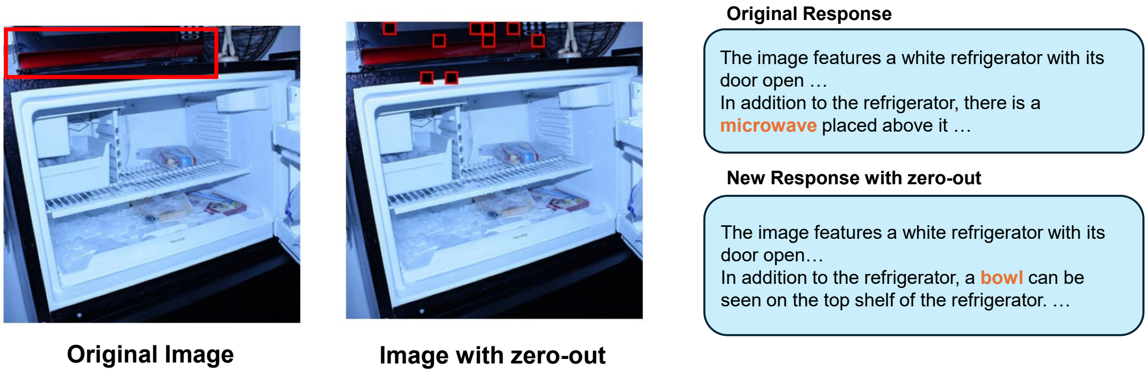
Original Response

The image features a young boy wearing a baseball helmet and holding a baseball bat, getting ready to swing at a ball. He is standing on a baseball field, surrounded by a sandy area. Another person can be seen …

New Response with zero-out

The image features a young boy wearing a baseball helmet and holding a baseball bat, getting ready to swing at a pitch. He is standing on a baseball field, fully focused on the game. Another person can be seen…

Figure 18. **Failure case of object “ball".** The original response successfully recognizes the “ball" on the ground. Due to the small size, applying zero-out directly masks the ball object from the image, resulting in the disappearance of the “ball" in the new response



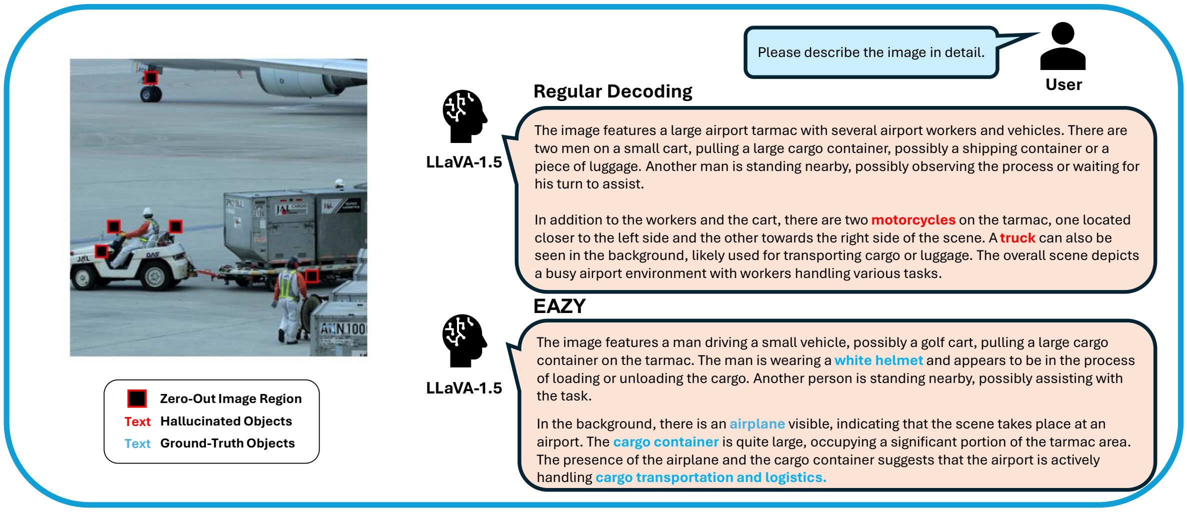
Original Response

The image features a white refrigerator with its door open …
In addition to the refrigerator, there is a microwave placed above it …

New Response with zero-out

The image features a white refrigerator with its door open…
In addition to the refrigerator, a bowl can be seen on the top shelf of the refrigerator. …

Figure 19. **Failure case of object “microwave".** The word “microwave" is split into “mic", “row" and “ave", which significantly increase the number of image tokens for zeroing out. In addition, this object is located on the edge of the image, with only part of it shown in the image. The new response does not have the “microwave" in it.



Please describe the image in detail.

Regular Decoding

The image features a large airport tarmac with several airport workers and vehicles. There are two men on a small cart, pulling a large cargo container, possibly a shipping container or a piece of luggage. Another man is standing nearby, possibly observing the process or waiting for his turn to assist.

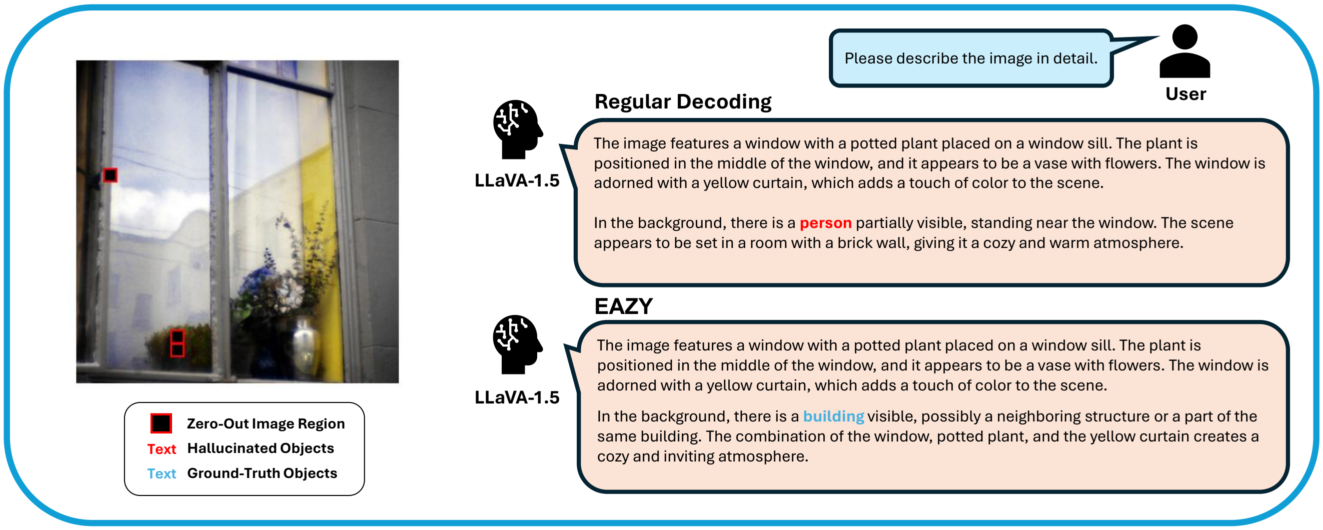
In addition to the workers and the cart, there are two motorcycles on the tarmac, one located closer to the left side and the other towards the right side of the scene. A truck can also be seen in the background, likely used for transporting cargo or luggage. The overall scene depicts a busy airport environment with workers handling various tasks.

EAZY

The image features a man driving a small vehicle, possibly a golf cart, pulling a large cargo container on the tarmac. The man is wearing a white helmet and appears to be in the process of loading or unloading the cargo. Another person is standing nearby, possibly assisting with the task.

In the background, there is an airplane visible, indicating that the scene takes place at an airport. The cargo container is quite large, occupying a significant portion of the tarmac area. The presence of the airplane and the cargo container suggests that the airport is actively handling cargo transportation and logistics.

Figure 20. **Case 1:** When using original greedy decoding, the response generated by LLaVA-1.5 model contains "motorcycles" and "truck", two HOs. After using EAZY to remove five image tokens, it can be observed the HOs disappeared in the new response with more details and better description.



Please describe the image in detail.

User

Regular Decoding

LLaVA-1.5

The image features a window with a potted plant placed on a window sill. The plant is positioned in the middle of the window, and it appears to be a vase with flowers. The window is adorned with a yellow curtain, which adds a touch of color to the scene.

In the background, there is a person partially visible, standing near the window. The scene appears to be set in a room with a brick wall, giving it a cozy and warm atmosphere.

EAZY

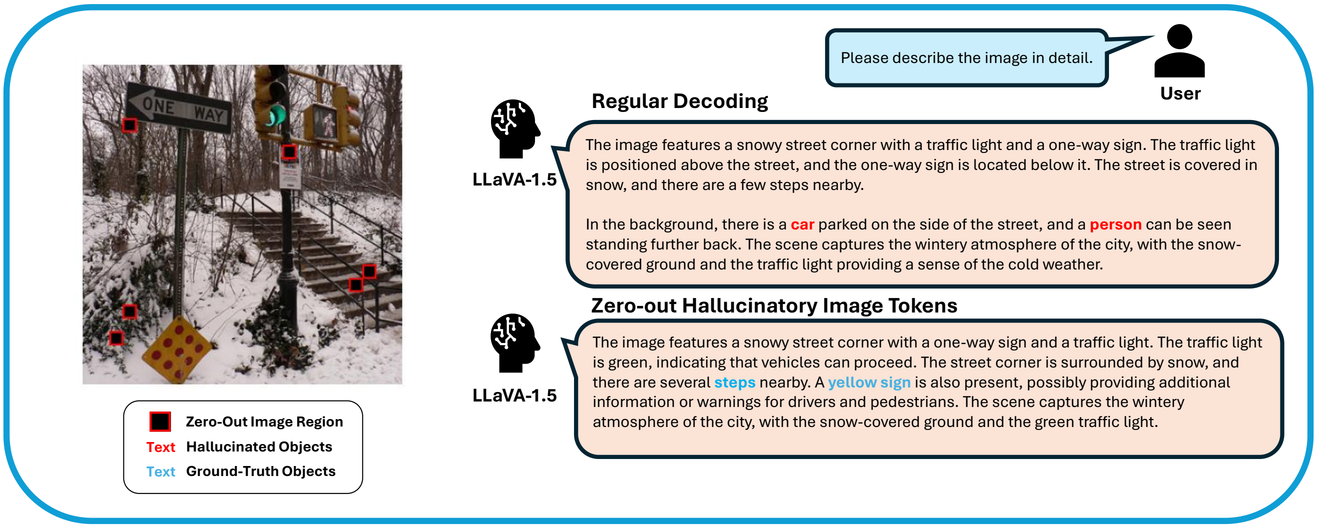
LLaVA-1.5

The image features a window with a potted plant placed on a window sill. The plant is positioned in the middle of the window, and it appears to be a vase with flowers. The window is adorned with a yellow curtain, which adds a touch of color to the scene.

In the background, there is a building visible, possibly a neighboring structure or a part of the same building. The combination of the window, potted plant, and the yellow curtain creates a cozy and inviting atmosphere.

■ Zero-Out Image Region
Text Hallucinated Objects
Text Ground-Truth Objects

Figure 21. **Case 2:** The LLaVA-1.5 falsely recognized a nonexistent person in the image. EAZY identified the building reflected in the window.



Please describe the image in detail.

User

Regular Decoding

LLaVA-1.5

The image features a snowy street corner with a traffic light and a one-way sign. The traffic light is positioned above the street, and the one-way sign is located below it. The street is covered in snow, and there are a few steps nearby.

In the background, there is a car parked on the side of the street, and a person can be seen standing further back. The scene captures the wintery atmosphere of the city, with the snow-covered ground and the traffic light providing a sense of the cold weather.

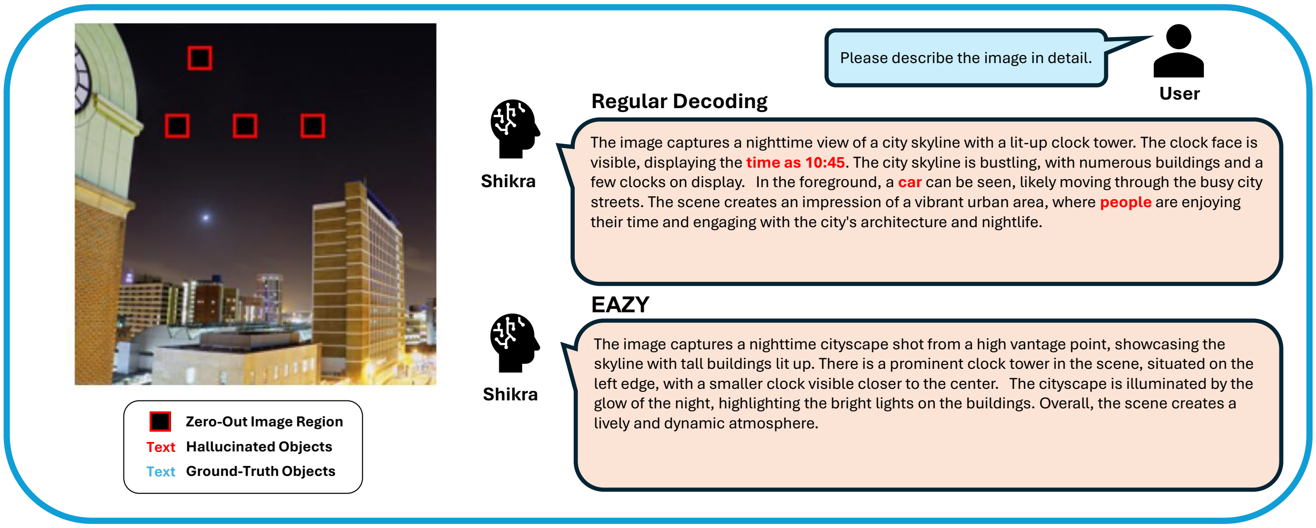
Zero-out Hallucinatory Image Tokens

LLaVA-1.5

The image features a snowy street corner with a one-way sign and a traffic light. The traffic light is green, indicating that vehicles can proceed. The street corner is surrounded by snow, and there are several steps nearby. A yellow sign is also present, possibly providing additional information or warnings for drivers and pedestrians. The scene captures the wintery atmosphere of the city, with the snow-covered ground and the green traffic light.

■ Zero-Out Image Region
Text Hallucinated Objects
Text Ground-Truth Objects

Figure 22. **Case 3:** In the original response, the model identified a car and a person in the image. EAZY successfully helped the model recognize the steps and the yellow markings while eliminating the previous hallucinations.



Please describe the image in detail.

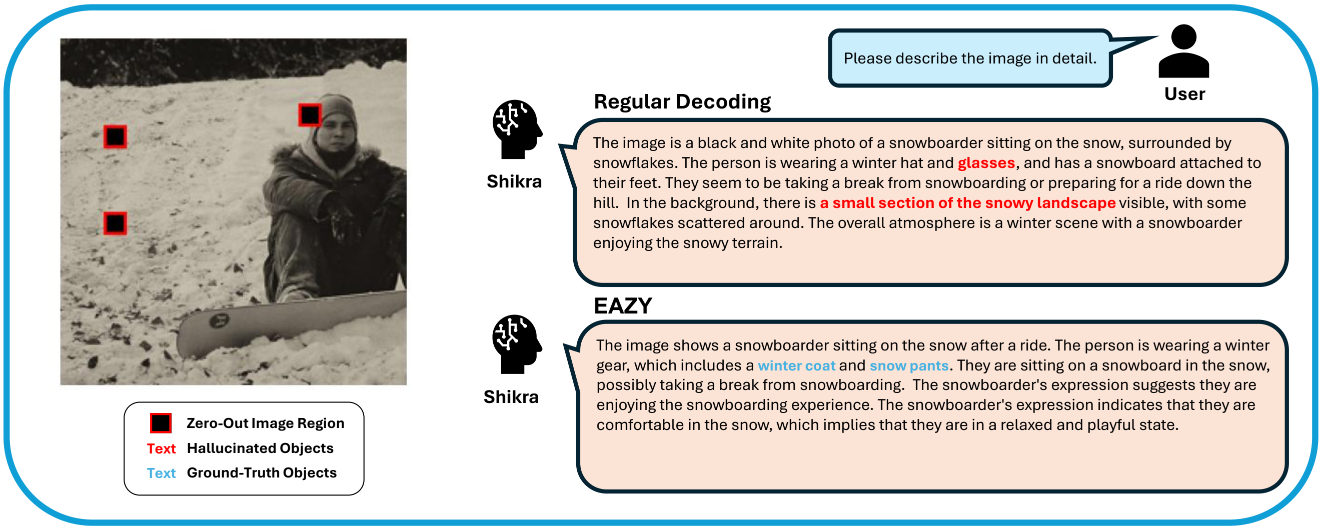
Regular Decoding

The image captures a nighttime view of a city skyline with a lit-up clock tower. The clock face is visible, displaying the **time as 10:45**. The city skyline is bustling, with numerous buildings and a few clocks on display. In the foreground, a **car** can be seen, likely moving through the busy city streets. The scene creates an impression of a vibrant urban area, where **people** are enjoying their time and engaging with the city's architecture and nightlife.

EAZY

The image captures a nighttime cityscape shot from a high vantage point, showcasing the skyline with tall buildings lit up. There is a prominent clock tower in the scene, situated on the left edge, with a smaller clock visible closer to the center. The cityscape is illuminated by the glow of the night, highlighting the bright lights on the buildings. Overall, the scene creates a lively and dynamic atmosphere.

Figure 23. **Case 4:** EAZY not only removed the hallucinated objects (*car* and *people*) but also corrected the inaccurate clock time.



Please describe the image in detail.

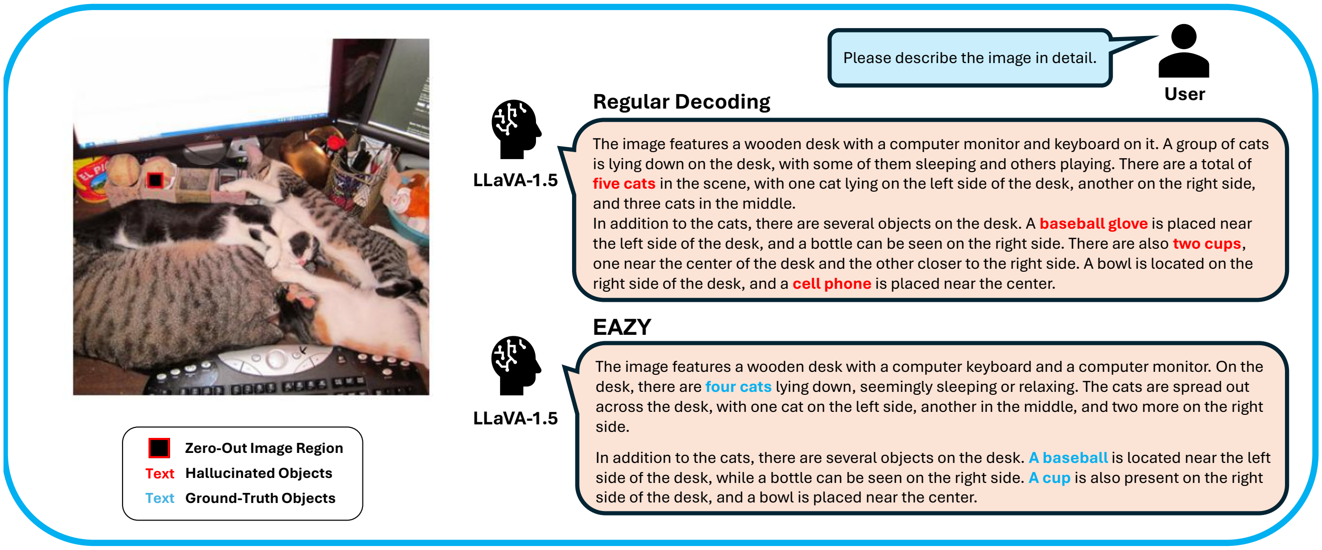
Regular Decoding

The image is a black and white photo of a snowboarder sitting on the snow, surrounded by snowflakes. The person is wearing a winter hat and **glasses**, and has a snowboard attached to their feet. They seem to be taking a break from snowboarding or preparing for a ride down the hill. In the background, there is **a small section of the snowy landscape** visible, with some snowflakes scattered around. The overall atmosphere is a winter scene with a snowboarder enjoying the snowy terrain.

EAZY

The image shows a snowboarder sitting on the snow after a ride. The person is wearing a winter gear, which includes a **winter coat** and **snow pants**. They are sitting on a snowboard in the snow, possibly taking a break from snowboarding. The snowboarder's expression suggests they are enjoying the snowboarding experience. The snowboarder's expression indicates that they are comfortable in the snow, which implies that they are in a relaxed and playful state.

Figure 24. **Case 5:** EAZY corrected the hallucinated object (*glasses*) and adjusted the inaccurate description of the background environment (*a small section of the snowy landscape*).



Please describe the image in detail.

Regular Decoding

The image features a wooden desk with a computer monitor and keyboard on it. A group of cats is lying down on the desk, with some of them sleeping and others playing. There are a total of **five cats** in the scene, with one cat lying on the left side of the desk, another on the right side, and three cats in the middle.

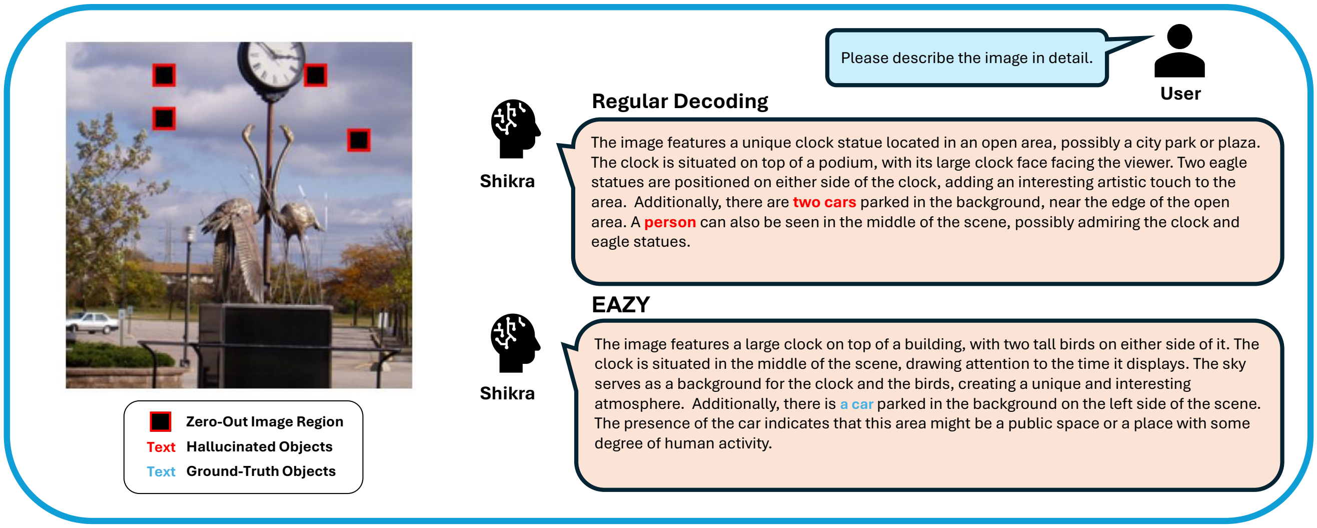
In addition to the cats, there are several objects on the desk. A **baseball glove** is placed near the left side of the desk, and a bottle can be seen on the right side. There are also **two cups**, one near the center of the desk and the other closer to the right side. A bowl is located on the right side of the desk, and a **cell phone** is placed near the center.

EAZY

The image features a wooden desk with a computer keyboard and a computer monitor. On the desk, there are **four cats** lying down, seemingly sleeping or relaxing. The cats are spread out across the desk, with one cat on the left side, another in the middle, and two more on the right side.

In addition to the cats, there are several objects on the desk. **A baseball** is located near the left side of the desk, while a bottle can be seen on the right side. **A cup** is also present on the right side of the desk, and a bowl is placed near the center.

Figure 25. **Case 6:** By zeroing out just a single image token, the model successfully identified the correct number of cats and cups in the image while removing the hallucinated objects *baseball glove* and *cell phone*.



Please describe the image in detail.

Regular Decoding

The image features a unique clock statue located in an open area, possibly a city park or plaza. The clock is situated on top of a podium, with its large clock face facing the viewer. Two eagle statues are positioned on either side of the clock, adding an interesting artistic touch to the area. Additionally, there are **two cars** parked in the background, near the edge of the open area. A **person** can also be seen in the middle of the scene, possibly admiring the clock and eagle statues.

EAZY

The image features a large clock on top of a building, with two tall birds on either side of it. The clock is situated in the middle of the scene, drawing attention to the time it displays. The sky serves as a background for the clock and the birds, creating a unique and interesting atmosphere. Additionally, there is a car parked in the background on the left side of the scene. The presence of the car indicates that this area might be a public space or a place with some degree of human activity.

Figure 26. **Case 7:** With EAZY, the number of cars in the new response was correctly identified, and the hallucinated object *person* was successfully removed.