

Same Attention, Different Truths: Put Logit-Lens over Visual Attention to Detect and Mitigate LVLMM Object Hallucination

Supplementary Material

A. Same Attention Across Models: Expanded Analysis and Visualizations

In the main paper (Sec.3.2), we identified the “Same Attention” phenomenon using LLaVA-1.5-7B, observing that real and hallucinated objects exhibit nearly identical attention magnitudes during the Image-Attention Stage. To validate the generalizability of this finding, we extend our analysis to three additional Large Vision-Language Models (LVLMMs): LLaVA-1.5-13B, Shikra-7B, and Qwen2-VL-7B. This section provides a comprehensive quantitative comparison and qualitative visualizations.

A.1. Layer-wise Attention Dynamics

We first analyze the evolution of attention weights assigned to image across layers for tokens generated by each model, as shown in Fig. 1. For object tokens, all models exhibit a distinct stage-wise structure containing a clear Image-Attention Stage:

- **LLaVA-1.5-7B/13B:** Consistent with the analysis in Sec.3.2, image attention follows a four-stage pattern: high in the initial 1–2 layers, dropping to a relatively low level in the early-mid layers, rising again to peak and sustaining the highest levels in the mid-to-late layers and finally dropping in the last few layers as the model focuses on language organization. Specifically, the Image-Attention Stage is located approximately at layers 20–27 for LLaVA-1.5-7B and layers 22–29 for LLaVA-1.5-13B.
- **Shikra-7B:** It exhibits a higher overall image attention than other models and still follows a four-stage pattern: lower in the first few layers, rising in the early-mid layers, further increasing to a peak in the mid-to-late layers, and declining in the final layers. Its Image-Attention Stage is identified approximately between layers 19–29.
- **Qwen2-VL-7B:** The attention evolution here closer resembles a three-stage structure: low in the early layers, a slight rise in the early-mid layers and reaching a maximum while sustaining high values in the mid-to-late layers. The corresponding Image-Attention Stage is approximately layers 15–28.

Despite differences in layer depth and curve shape, all models demonstrate a distinct Image-Attention Stage,

where attention to the image is significantly higher than in other phases and remains stable. Furthermore, the attention curves for real object tokens and hallucinated object tokens are highly similar across all four models. They not only share comparable magnitudes but also exhibit synchronized Image-Attention Stages. This suggests that the amount of attention a model pays to the image—how much it attends to the image, does not explain why hallucination occurs.

In contrast, the curves for non-object tokens are much flatter. Their attention to the image is consistently lower than that of object tokens across all layers, and they lack an Image-Attention Stage.

A.2. Statistical Comparison Within the Image-Attention Stage

After identifying the Image-Attention Stage for each model, we calculate the mean and standard deviation of the image attention for all generated tokens within their respective Image-Attention Stages, and aggregate the results by category (Real Object Tokens, Hallucinated Object Tokens and Non-Object Tokens). As shown in Fig. 2:

- There is no statistically significant difference in attention magnitude between real object tokens and hallucinated object tokens for each model.
- The visual attention for both object categories is significantly higher than that for non-object tokens.

Although the absolute attention ratios vary across architectures (e.g., Shikra-7B shows generally higher image attention), the relative trend remains identical. This further quantitatively validates the core conclusion of the main paper: during the Image-Attention Stage, where visual processing is most critical, hallucinated objects do not receive significantly less visual attention. In other words, “insufficient visual attention” cannot explain the phenomenon of object hallucination.

A.3. Layer-Wise Attention Heatmaps

To provide a more intuitive understanding of these statistical findings, we present more layer-wise attention heatmaps in Fig. 3 comparing real and hallucinated object tokens. These visualizations demonstrate that for both real and hallucinated objects, attention gradually converges to specific

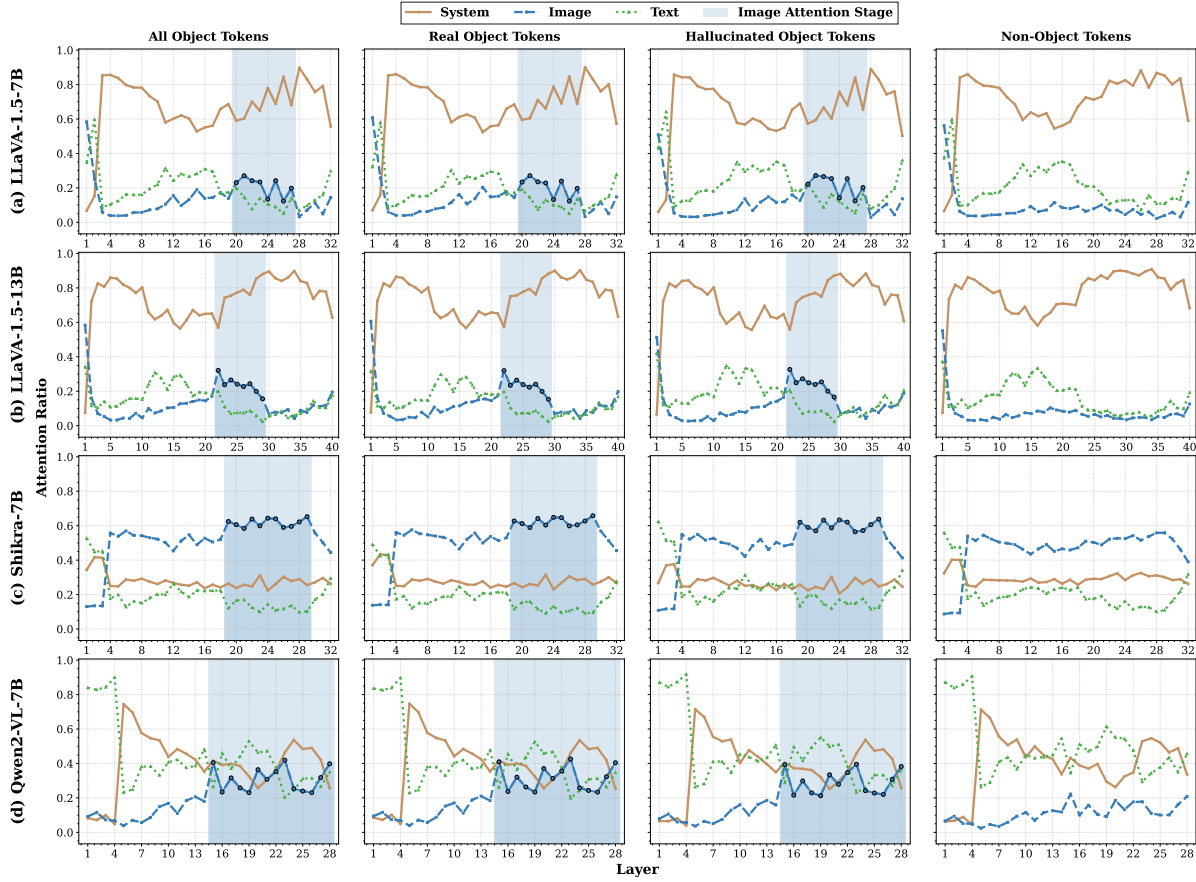


Figure 1. Layer-wise attention of different token types across different models. The highlighted region denotes the Image-Attention Stage

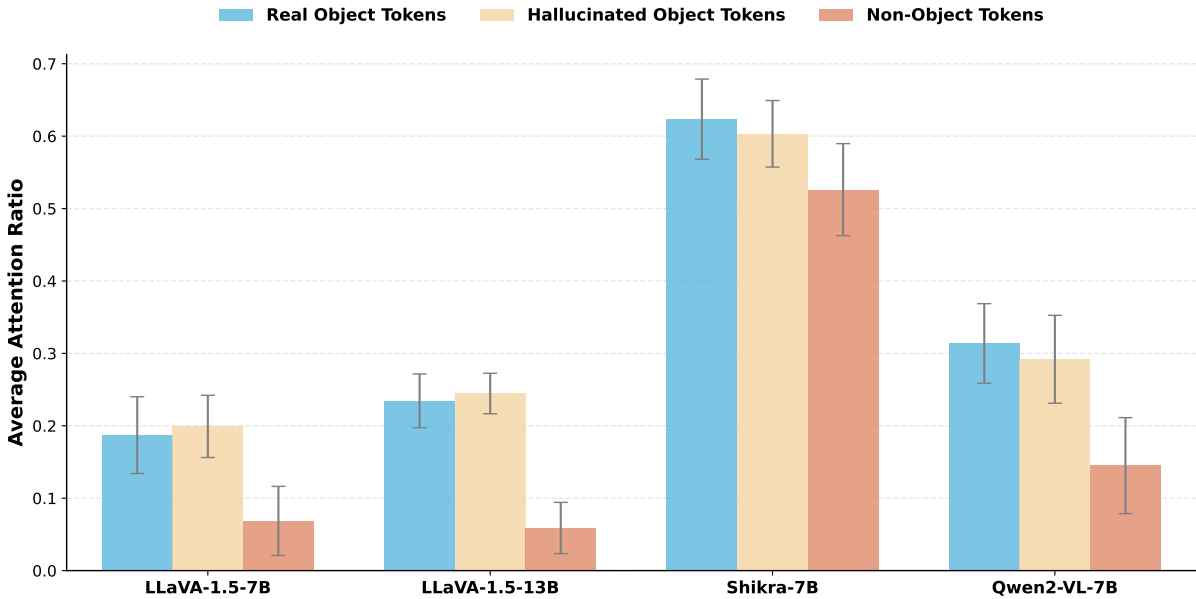


Figure 2. Comparison of the average image attention within the Image-Attention Stage for real object, hallucinated object, and non object tokens across different models.

image regions as layers progress. Crucially, within the Image-Attention Stage of each model, the heatmaps display concentrated and relatively stable highlighted regions. These visual results corroborate our argument: hallucination occurs when the interpretation of the high attention regions (what it attends to) is flawed (leading to Visual Uncertainty) or is overridden by language priors (Contextual Prior).

B. Different Truths Across Models: Extended Qualitative Visualizations

To further corroborate the “Different Truths” phenomenon discussed in Sec.3.3 of the main paper, we present extended qualitative visualizations based on Logit-Lens across different models in Fig. 4. These results provide intuitive evidence of the significant disparity in the relationship between visual semantic decoding and generated tokens for real versus hallucinated objects.

In the figure, odd-numbered rows correspond to real object tokens. When the model generates an object that is actually present, the words decoded by Logit Lens from high-attention regions are highly consistent with the final output tokens (including synonyms and subword pieces). This indicates that the model not only “looks at” the correct regions, but also correctly understands its visual semantics. In contrast, for the even rows (hallucinated object tokens), although the model exhibits attention intensity comparable to that of real objects (i.e., Same Attention), Logit-Lens reveals a severe semantic inconsistency between the underlying visual semantics and the generated text.

C. Two Types: Extended Cases

In this section, we present richer qualitative examples to further illustrate the two types of hallucination identified in the main paper, as shown in Fig. 5.

C.1. Type 1: Visual Uncertainty Hallucination

The core characteristic of this type is that while the model’s attention is indeed concentrated on specific image regions, these regions are visually ambiguous, incomplete, or possess only weak associations with the target category, failing to provide stable semantic evidence. When these “high-attention but unreliable” regions are masked, the model immediately ceases to generate the hallucinated object. This indicates that the hallucination stems directly from a failure in visual understanding. For instance, in the case of the hallucinated “handbag”, attention focuses on clothing folds or small shadow patches; for the hallucinated “bicycle”, attention falls on distant grass or blurry textures. Notably, Logit-Lens reveals that the tokens decoded from these regions are also inconsistent with the actual objects in the image. For example, the region attended to for “handbag” is decoded as

“bottle”, and the region for “bicycle” is decoded as “truck” or “motor”.

C.2. Type 2: Contextual Prior Hallucination

The core characteristic of this type is that the generation of an object is unrelated to the attended visual region and is instead driven by strong linguistic or scene co-occurrence priors. The model’s attention behavior appears “formalistic”—acting as if it “must look somewhere before generating” to satisfy an internal procedural requirement, such as generating a “microwave” in a kitchen description or a “chair” in a bedroom description. Even when the high-attention regions are masked, the hallucination persists, with attention often drifting to other regions. Crucially, unlike Type 1, the tokens decoded from the high-attention regions in these cases are often consistent with the actual objects present in those regions. For example, when hallucinating a microwave, the model attends to a region decoded as “vent” (ventilation); when hallucinating a chair, it attends to a region decoded as “lamp”. This observation provides the theoretical underpinning for our VEED method: we can leverage this correct visual evidence (captured by Logit-Lens) to suppress the erroneous priors.

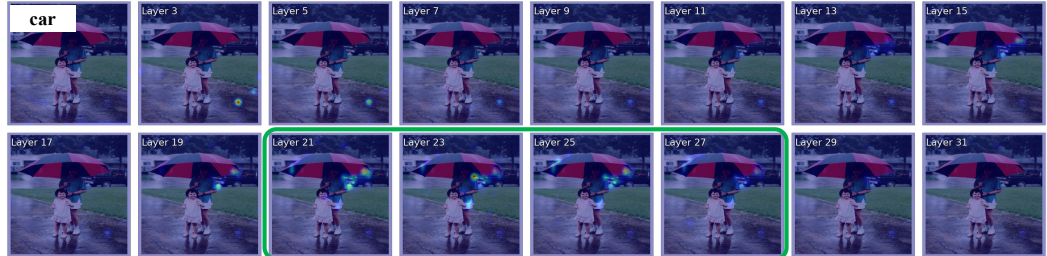
D. Extended Analysis of Hallucination Detection

This section provides additional implementation details and quantitative analyses for the LLCC proposed in the main paper.

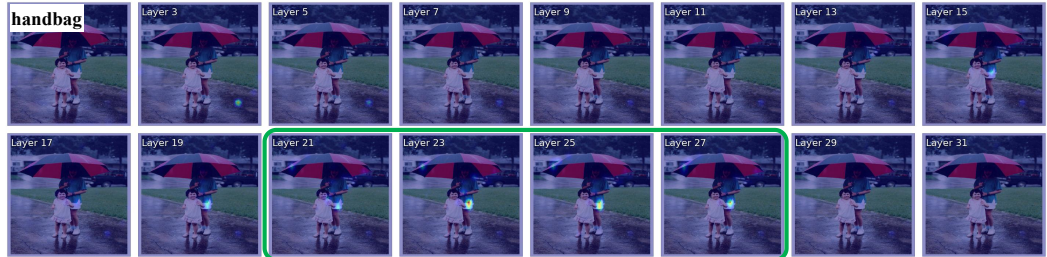
D.1. More Details

In Sec.4.1.1 of the main paper, we proposed filtering object tokens based on the average attention score $\mathcal{A}_{img}(o_t)$ within the Image-Attention Stage. However, empirical observations reveals that certain non-object tokens—specifically adjectives describing attributes (e.g., red, large) or quantifiers (e.g., two)—can exhibit relatively higher attention magnitudes compared to semantically meaningless function words (e.g., prepositions), even though they generally remain lower than genuine object nouns. To eliminate potential interference from these descriptive terms during detection, we employ Part-of-Speech (POS) tagging as a secondary filter. We specifically retain only tokens tagged as nouns (e.g., NN, NNS), ensuring that the consistency check focuses exclusively on substantive object entities.

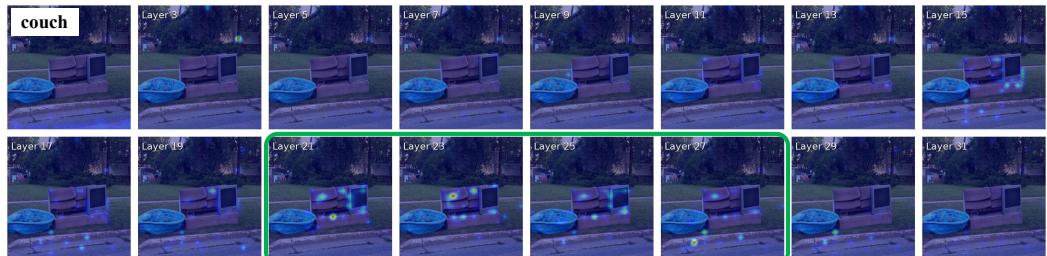
For semantic consistency checking, outputs decoded from Logit-Lens may be sub-word tokens. Therefore, we first construct a semantic similarity set for each target object name using the similarity threshold τ_{sim} and. A decoded token is then considered semantically consistent if it belongs to any word of this set.



Q: Describe this image.
A: In the background, there is a **car** parked on the right side of the scene. Additionally, there is a **handbag** near the woman and the girl, possibly belonging to one of them.



Q: Describe this image.
A: The image features a young child sitting on a **suitcase**, which is placed on the floor. A **book** can be seen on the floor, and a **cell phone** is located near the suitcase.



Q: Describe this image.
A: The image features a curbside scene with a **couch** and a television set placed on the grass. Additionally, there is a **bicycle** parked further back in the scene, closer to the left side.

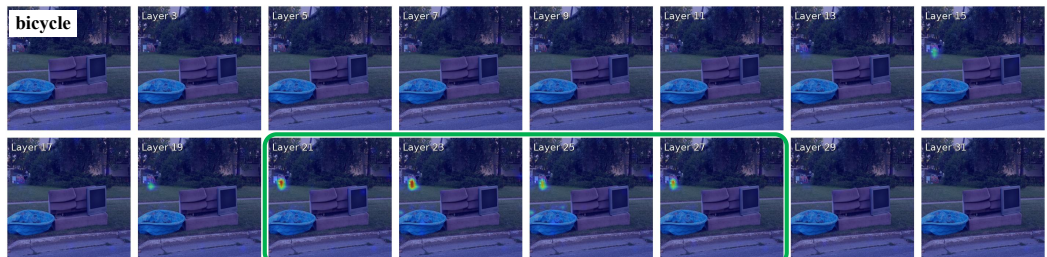


Figure 3. Layer-wise attention maps for real object and hallucinated object tokens. The green box indicates the Image-Attention Stage.



Figure 4. Logit-Lens decoding results for real and hallucinated tokens across different models.

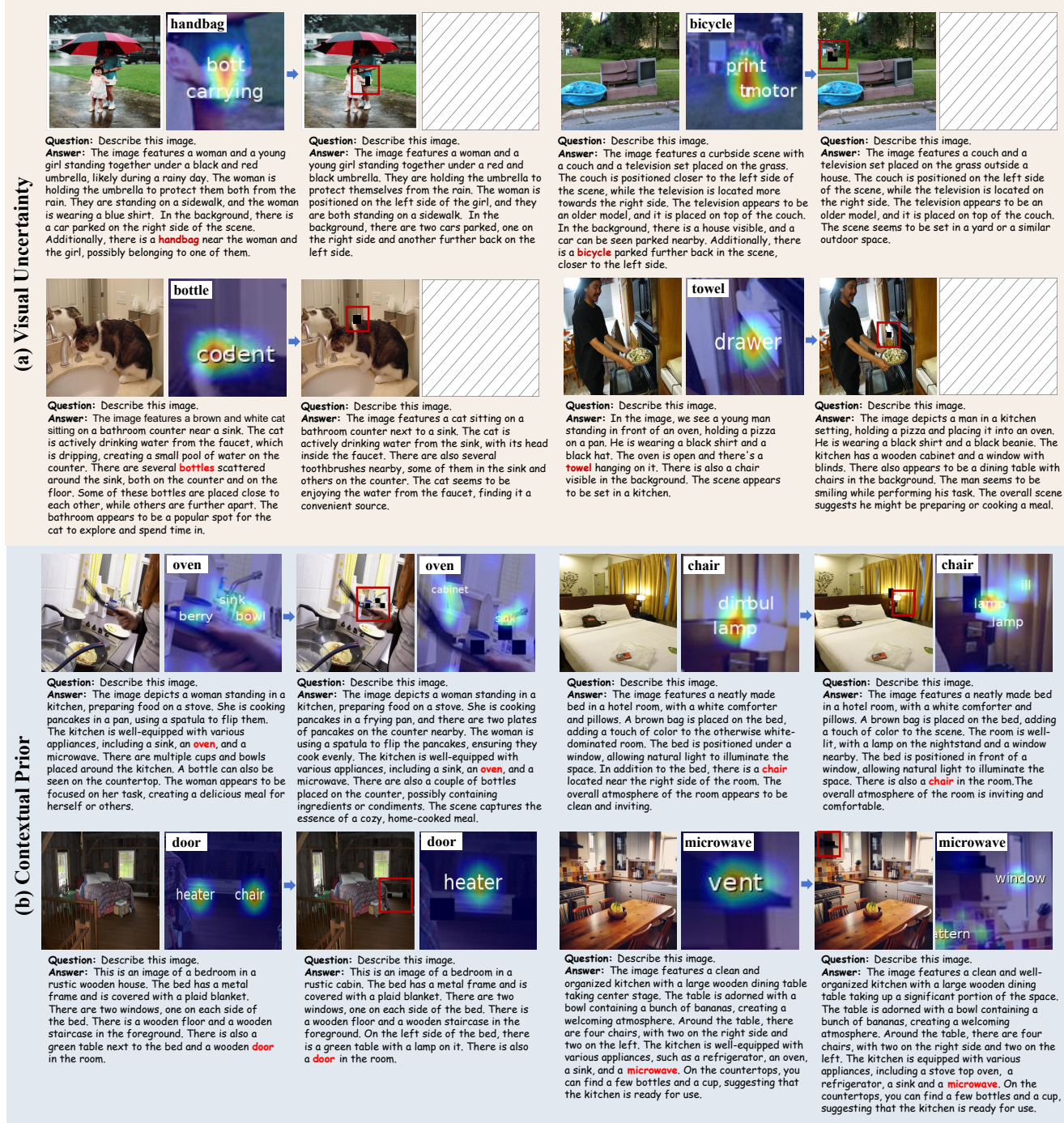


Figure 5. Additional examples for the two hallucination types.

D.2. Ablation Study on Top- k

We investigated the impact of the Logit-Lens decoding candidate number k on detection performance using LLaVA-1.5-7B. As shown in Tab. 1, we observe a clear trade-off: as k increases, precision gradually improves while recall de-

clines. The F1 score reaches its peak when $k = 3$. This trend is theoretically sound. When k is small (e.g., $k = 1$), the criterion is overly strict, requiring the model to decode the exact correct target token (or its synonym) from only a single attended region. This restricted scope leads to the over-identification of hallucinations (high recall, low pre-

cision), while a large k shows the opposite trend. $k = 3$ provides the best balance, and we adopt it as the default in all experiments.

Table 1. Top- k ablation study for hallucination detection. Best results are in **bold**.

Top- k	Precision	Recall	F1 Score
1	0.5884	0.8514	0.6959
2	0.7027	0.7995	0.7480
3	0.7870	0.7955	0.7932
4	0.8025	0.7230	0.7607
5	0.8115	0.6689	0.7333

D.3. Detection Performance Across Models

Tab. 2 presents the detection results across four different LLMs. LLaVA-1.5-7B achieves the best detection performance, while Qwen2-VL yields the lowest scores. This performance disparity is attributable to the specific internal representation characteristics of Qwen2-VL. In our experiments, we observe that Logit-Lens can only decode meaningful textual semantics in the final 1–2 layers for Qwen2-VL. In the intermediate layers, the decoded outputs often consist of non-standard or semantically unrelated tokens. Consequently, the effective range of layers available for our semantic consistency check is significantly more restricted in Qwen2-VL compared to LLaVA, resulting in comparatively lower detection accuracy.

Table 2. Detection results across different models. Best results are in **bold**.

Model	Precision	Recall	F1 Score
LLaVA-1.5-7B	0.7870	0.7955	0.7932
LLaVA-1.5-13B	0.7543	0.7721	0.7631
Shikra-7B	0.7178	0.7210	0.7194
Qwen2-VL-7B	0.6898	0.7103	0.6999

E. Ablation Studies on Mitigation Strategy

To thoroughly evaluate the contributions of each component within our Detect-Mitigate framework, we conducted ablation studies on LLaVA-1.5-7B. Our analysis focuses on two primary aspects: (1) the effectiveness of the two mitigation strategies (HARM and VEED) designed for different hallucination types; and (2) the impact of the visual evidence enhanced intensity, α , within the VEED strategy.

E.1. Effectiveness of Mitigation Components

Our framework designs specific remedies for two distinct hallucination mechanisms: HARM for Type 1 (Visual Uncertainty) and VEED for Type 2 (Contextual Prior). To verify the necessity of this targeted treatment, we tested the performance when applying each strategy individually versus jointly, as shown in Tab. 3.

- Only HARM: When masking is applied solely for all the hallucinated tokens, the $CHAIR_S$ and $CHAIR_I$ drops from 49.8 to 34.2 and 20.4 to 13.9, respectively. This indicates that masking visual evidence resolves a portion of the errors, yet hallucination driven by contextual priors remain prevalent.
- Only VEED: When only Visual Evidence Enhanced Decoding is applied to all hallucinated tokens, it shows a smaller improvement. This suggests that without addressing Type 1 errors, the model continues to be misled by erroneous visual features. Merely strengthening the visual signal is insufficient to correct these perceptual errors.
- Joint Application (Full Method): When both strategies are enabled simultaneously, performance improves significantly. This demonstrates that the two hallucination mechanisms are distinct yet co-existing, and that effective mitigation requires “treating the right cause”: cutting off unreliable visual sources for visual uncertainty cases while injecting correct visual evidence when contextual priors dominate.

Table 3. Ablation study of the mitigation methods. Best results are in **bold**.

HARM	VEED	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$
		49.8	20.4
✓		34.2	13.9
	✓	40.8	16.4
✓	✓	26.8	10.0

E.2. Experiment on α

For the second type of hallucination (Contextual Prior), we introduce a hyperparameter α to control the intensity of visual evidence injection. The results are presented in Tab. 4. We observe that the mitigation effectiveness declines as α increases. The model achieves the lowest hallucination rate when $\alpha = 0$, which corresponds to fully replacing the original logits with the visual logits derived from Logit-Lens. This suggests that, the original predictions contain very little correct information, and relying entirely on visual evidence is the most effective way to suppress the contextual prior.

Table 4. Ablation study on α . Best results are in **bold**.

α	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$
0	26.8	10.0
0.1	27.9	10.8
0.2	29.5	11.9
0.3	33.1	13.2

F. Limitations and Future Work

While this work provides a systematic analysis of LVLM hallucination mechanisms through the lens of attention patterns and semantic decoding, and proposes a training-free Detect-Mitigate framework, limitations still remain. First, our consistency check relies on analyzing the semantics of high-attention regions during the Image-Attention Stage. However, some models exhibit the “attention sink” phenomenon, where excessive attention is directed towards non-informative tokens. Although methods exist to identify and filter these sink tokens, their presence may still introduce noise to our attention-based localization. Second, some models exhibit limited semantic decodability in intermediate layers when using Logit Lens, so we plan to explore learnable probing methods to more robustly evaluate the semantic content encoded in these representations. We also plan to conduct deeper investigations into the underlying mechanisms of the two hallucination types, aiming to more fundamentally understand their causal origins. Finally, while this work focuses on object hallucination, future research will extend to other forms of hallucination, including attribute hallucination, relation hallucination, and reasoning hallucination.

G. Case Study

We present several mitigation results produced by our method in the following.



Question: Describe this image.

Original Answer: The image features a bathroom with a large **mirror** on the wall, reflecting a **television** mounted above the sink. The television is displaying a football game, capturing the attention of the viewers. In the bathroom, there are two **sinks**, one on the left side and another on the right side. Above the sinks, there are two **toothbrushes** placed, likely belonging to the users of the bathroom. Additionally, there is a **cup** placed on the right side of the bathroom, possibly for holding toothbrushes or other bathroom essentials.

Our Answer: The image features a bathroom with a large **mirror** on the wall, reflecting a **television screen**. The television is mounted on the wall, and the reflection of the screen is clearly visible in the mirror. The bathroom appears to be well-maintained and clean. In the reflection, there are two **people** visible, one on the left side and another on the right side of the screen. The bathroom also has two **sinks**, with one located on the right side and the other on the left.



Question: Describe this image.

Original Answer: The image features a man and a young boy sitting on a bed, enjoying a **pizza** together. The man is holding a slice of pizza in his hand, and the boy is eagerly waiting for his share. The pizza is placed on a **dining table** in the room, which also has a **chair** and a **bottle** on it. In addition to the main subjects, there are a few other objects in the room, including a **teddy bear**, a **remote control**, and a **cup**. The scene appears to be a cozy and fun moment shared between the man and the child.

Our Answer: The image features a man and a young boy sitting on a bed, enjoying a **pizza** together. The man is holding a slice of pizza in his hand, and the boy is eagerly waiting for his share. The pizza is placed on a **box**, which is placed on the bed. There is also a **bottle** located near the center of the scene. The scene appears to be a cozy and fun moment between the man and the child.



Question: Describe this image.

Original Answer: The image features a large, fluffy white **dog** lying on a **couch**, curled up and resting comfortably. The dog is positioned in the center of the couch, occupying a significant portion of the space. In the background, there is a **dining table** with a **chair** placed near it. Additionally, there is a **vase** placed on the table, possibly as a decorative element.

Our Answer: The image features a large, fluffy white **dog** lying on a **couch**, curled up and resting comfortably. The dog is positioned in the center of the couch, occupying a significant portion of the space. In the background, there is a **chair** located near the left side of the room.



Question: Describe this image.

Original Answer: The image features a wooden **deck** with a red **pot** filled with green **plants**, placed in the center of the deck. The pot is surrounded by a wooden **fence**, and there is a **bench** located nearby. A large **umbrella** is open and covering the entire deck, providing shade and protection from the elements. In the background, there is a **chair** positioned on the deck, and a **dining table** can be seen further away. The scene appears to be a cozy outdoor space, perfect for relaxation and enjoying the greenery.

Our Answer: The image features a wooden **deck** with a red **pot** filled with green **plants**, placed in the center of the deck. The pot is surrounded by a wooden **fence**, and there is a striped **umbrella** positioned above the pot, providing shade and protection from the elements. In addition to the main plant, there are a **chair** placed on the right side of the deck. The scene appears to be set in a backyard or outdoor space, with a striped umbrella providing a cozy atmosphere.

Figure 6. The mitigation performance of our method.



Question: Describe this image.

Original Answer: The image features a neatly made **bed** in a hotel room, with a white **comforter** and **pillows**. A brown **bag** is placed on the bed, adding a touch of color to the otherwise white-dominated room. The bed is positioned under a window, allowing natural light to illuminate the space. In addition to the bed, there is a **chair** located near the right side of the room. The overall atmosphere of the room appears to be clean and inviting.

Our Answer: The image features a neatly made **bed** in a hotel room, with a white **comforter** and **pillows**. A brown **bag** is placed on the bed, adding a touch of color to the otherwise white-dominated room. The bed is positioned under a window, allowing natural light to illuminate the space. In addition to the bed, there is a **lampshade** on the nightstand. A **handbag** is also visible in the room, placed near the bed.



Question: Describe this image.

Original Answer: The image depicts a busy street scene with a man riding a **bicycle** down the road, carrying a large number of **chickens** in a cart. The chickens are spread across the cart, with some closer to the front and others further back. In addition to the man on the bicycle, there are several other people in the scene, some walking and others standing. A **car** is visible on the right side of the image, and a **truck** can be seen further back on the left side. There are also a couple of **umbrellas** in the scene, one near the left edge of the image and another towards the center. A **handbag** is placed on the ground near the left side of the image, and a **potted plant** is located on the right side of the scene.

Our Answer: The image depicts a busy street scene with a man riding a **bicycle** down the road, carrying a large number of **chickens** in a cart. The chickens are spread across the cart, with some closer to the front and others further back. In addition to the man on the bicycle, there are several other people in the scene, some walking and others standing. A **car** is visible on the right side of the image, and a **hat** can be seen on one of the people in the scene. The overall atmosphere suggests a lively and bustling street environment.



Question: Describe this image.

Original Answer: The image features a group of people sitting around a dining table, enjoying a meal together. There are several **cups** placed on the table, with some of them containing water. A **cell phone** is also visible on the table, likely belonging to one of the diners. The table is set with various utensils, including **forks**, **knives**, and **spoons**. There are also a couple of **bowls** on the table, possibly containing food or condiments. The atmosphere appears to be casual and social, with everyone engaged in conversation and enjoying their time together.

Our Answer: The image features a group of people sitting around a dining table, enjoying a meal together. There are several **cups** placed on the table, with some of them containing water. A dessert **menu** is also visible on the table, likely indicating the dessert options available for the guests. In addition to the cups, there are multiple **bottles** scattered around the table, possibly containing beverages or condiments. A **cell phone** can be seen on the table, possibly belonging to one of the guests. The atmosphere appears to be relaxed and social, with everyone gathered around the table to share a meal and conversation.



Question: Describe this image.

Original Answer: The image features a clean and organized kitchen with a large wooden dining table taking center stage. The table is adorned with a **bowl** containing a bunch of **bananas**, creating a welcoming atmosphere. Around the table, there are four **chairs**, with two on the right side and two on the left. The kitchen is well-equipped with various appliances, such as a **refrigerator**, an **oven**, a **sink**, and a **microwave**. On the countertops, you can find a few bottles and a cup, suggesting that the kitchen is ready for use.

Our Answer: The image features a clean and organized kitchen with a large wooden dining table taking center stage. The table is adorned with a **bowl** containing a bunch of **bananas**, creating a welcoming atmosphere. Around the table, there are four **chairs**, with two on the right side and two on the left. The kitchen is well-equipped with various appliances, such as a **refrigerator**, an **oven**, a **sink**, and a **ventilation** system, which provides air circulation to the space. The overall scene is a pleasant and functional space for dining and socializing.

Figure 7. The mitigation performance of our method.



Question: Describe this image.

Original Answer: The image features a large, delicious-looking **hot dog sandwich** sitting on a table. The hot dog is placed on a bun and is accompanied by a variety of toppings, including mustard and ketchup. The sandwich is accompanied by a side of **fries**, which are placed in a **bowl** nearby. In addition to the main dish, there is a small American **flag** placed on top of the hot dog, adding a patriotic touch to the meal. The table setting is complete with a **fork** and a **knife**, which are placed on the table, ready for use.

Our Answer: The image features a large, delicious-looking **hot dog sandwich** sitting on a table. The hot dog is placed on a bun and is accompanied by a variety of toppings, including mustard and ketchup. The sandwich is quite tall, making it an impressive and appetizing meal. In addition to the hot dog, there is a **bowl** of **fries** nearby, adding to the meal's appeal. A **stick** with an American **flag** on it is also present, possibly as a decoration or a toothpick for the hot dog. The overall scene is inviting and ready to be enjoyed.



Question: Describe this image.

Original Answer: The image features a red double-decker **bus** driving down a street, with a group of people riding on the top deck. There are at least 13 **people** visible on the bus, enjoying the ride and the view from the open-top bus. Some of the **passengers** are standing, while others are sitting, all of them seemingly having a good time. In addition to the bus, there are a few other vehicles in the scene, including a **car** and a **truck**. A couple of **traffic lights** can also be seen, indicating that the bus is driving in an urban environment.

Our Answer: The image features a red double-decker **bus** driving down a street, with a group of people riding on the top deck. There are at least 13 **people** visible on the bus, enjoying the ride and the view from the open-top bus. Some of the **passengers** are standing, while others are sitting, all of them seemingly having a good time. The bus is quite large, occupying a significant portion of the image, and it appears to be a popular mode of transportation for the people in the scene. The passengers are spread out across the top deck, with some closer to the front and others near the back of the bus.



Question: Describe this image.

Original Answer: The image depicts a cozy living room with a white **couch** and a **television** placed on a wooden stand. The room is furnished with a coffee **table** and a few **potted plants**, creating a warm and inviting atmosphere. There are two couches in the room, one on the left side and another on the right side. A **chair** is also present in the room, situated near the left couch. The television is positioned in the middle of the room, with a **remote control** placed nearby. In addition to the main furniture, there are several decorative items in the room, including a **vase**, a potted plant, and a **bowl**. A **book** can also be seen on the coffee table, adding to the room's lived-in feel.

Our Answer: The image depicts a cozy living room with a white **couch** and a **television** placed on a wooden stand. The room is furnished with a coffee **table** and a few **potted plants**, creating a warm and inviting atmosphere. There are two potted plants in the room, one located near the couch and the other near the television. A **vase** can also be seen in the room, adding to the room's decor. The television is positioned on the right side of the room, while the couch is situated on the left side. The living room appears to be well-organized and comfortable, making it an ideal space for relaxation and entertainment.



Question: Describe this image.

Original Answer: The image features a yellow **car** parked in a parking lot near a building. The car is positioned in the middle of the scene, and it appears to be an older model. The parking lot is surrounded by a few **boats**, with some of them located closer to the right side of the image. In addition to the car and boats, there are two **benches** visible in the scene. One bench is situated near the left side of the image, while the other is located closer to the right side. The presence of these benches suggests that the parking lot may be a popular spot for people to relax and enjoy the view of the boats.

Our Answer: The image features a yellow **car** parked in a parking lot near a building. The car is positioned in the middle of the scene, and it appears to be an older model. The parking lot is surrounded by a few **boats**, with some of them located closer to the right side of the image. In addition to the car and boats, there are two **Hinweis signs** visible in the scene. One is located near the right edge of the image, while the other is situated closer to the center. The presence of these signs suggests that the parking lot might be near a body of water or a marina.

Figure 8. The mitigation performance of our method.