Devils in Middle Layers of Large Vision-Language Models: Interpreting, Detecting and Mitigating Object Hallucinations via Attention Lens

Supplementary Material

A. Limitations

Despite the simplicity and effectiveness of our hallucination detection and mitigation, there are several limitations:

- First, the SVAR metric, used to detect hallucinated object tokens, is limited by the inherent attention behavior of LVLMs. When LVLM consistently exhibits extremely high visual attention ratios at nearly all layers, such as the case of Shikra illustrated in Fig. 13 (a), this may weaken the effectiveness of the SVAR metric.
- Second, although the use of the VAR score and logit lens approach can intuitively distinguish two stages of visual information processing, identifying the specific range of these stages remains somewhat subjective. However, leveraging learnable strategies, such as training a set of learnable weights for layers based on the signals from VAR distribution and prediction contributions, could potentially achieve automatic localization of these stages, and we leave this for future work.

B. Experiment Details

B.1. Datasets for Case Study

Tab. 6 reports the statistical information of the synthetic datasets used in our case studies. Additionally, Fig. 9 illustrates the positional distributions of real and hallucinated object tokens for the four selected LVLMs.

B.2. MLP Training Details

Fig. 10 shows the training pipeline of the object hallucination detector. Tab. 7 details the hyperparameters used to train the two-layer MLP, designed for detecting hallucinated object tokens as described in Sec. 3.4.1. The Adam optimizer is employed to train the classifier with the number of epochs set to 200. For each layer range, we utilize a gride search strategy to find the optimal hidden layer size and learning rate within the ranges of {64, 128, 256, 512} and {1e-2, 1e-3, 1e-4}, respectively.

Model	No. of Real	No. of Hallucinated
LLaVA-1.5-7B	4,397	1,842
LLaVA-1.5-13B	4,488	1,700
Shikra-7B	4,263	1,794
MiniGPT-4-7B	2,999	981

Table 6. Statistical information of case datasets.

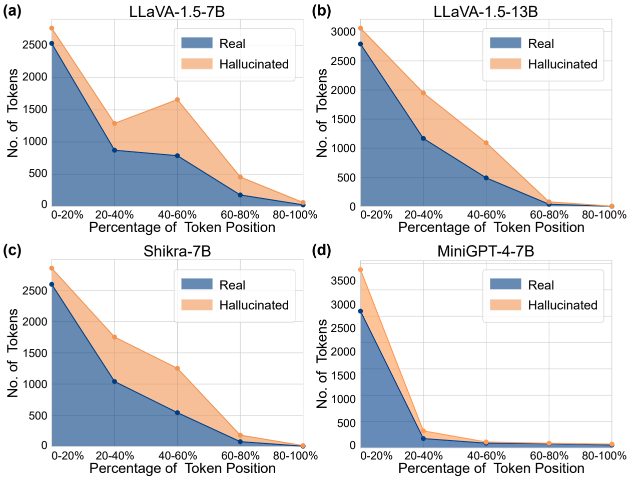


Figure 9. Real and hallucinated object token distributions by their position in description (%).

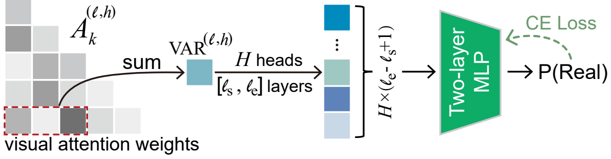


Figure 10. Illustration of detecting hallucinated object tokens by training an MLP classifier on the concatenated VAR scores.

Hyperparameters	LLaVA-1.5-7B
Optimizer	Adam [22]
(β_1, β_2)	(0.9, 0.999)
Hidden size	{64, 128, 256, 512}
Learning rate	{1e-2, 1e-3, 1e-4}
No. of epochs	200

Table 7. Training hyperparameters of the two-layer MLP for hallucination detection on LLaVA-1.5-7B.

C. Additional Results

C.1. Case Study Results

In this subsection, we conduct additional experiments on LLaVA-1.5-13B, Shikra-7B, and MiniGPT-4-7B to examine whether other models also share similar characteristics with LLaVA-1.5-7B.

LLaVA-1.5-13B. Fig. 11 (a) and (b) show the VAR score distribution and the prediction contributions from the

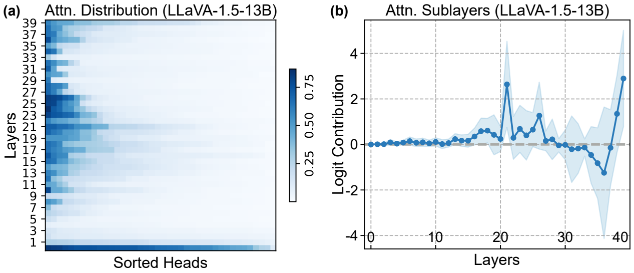


Figure 11. (a) Distribution of visual attention ratio for real object tokens across heads and layers in LLaVA-1.5-13B, sorted row-wise by attention ratios. (b) The logit contribution of attention sublayers to real token prediction.

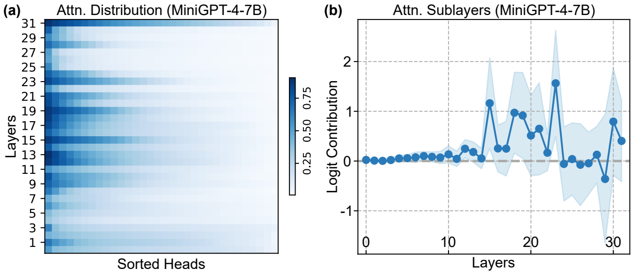


Figure 12. (a) Distribution of visual attention ratio for real object tokens across heads and layers in MiniGPT-4-7B, sorted row-wise by attention ratios. (b) The logit contribution of attention sublayers to real token prediction.

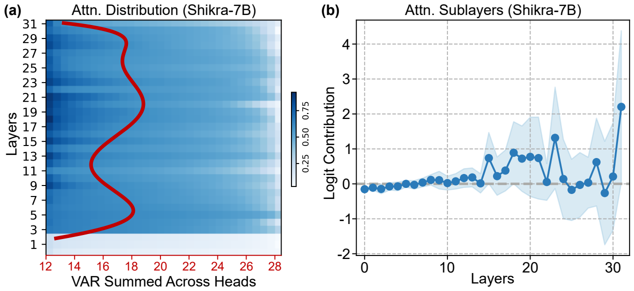


Figure 13. (a) Distribution of visual attention ratio for real object tokens across heads and layers in Shikra-7B, sorted row-wise by attention ratios. Note that the red curve represents a seventh-order polynomial fit to the values of attention ratios summed over heads in each layer. (b) The logit contribution of attention sublayers to real token prediction.

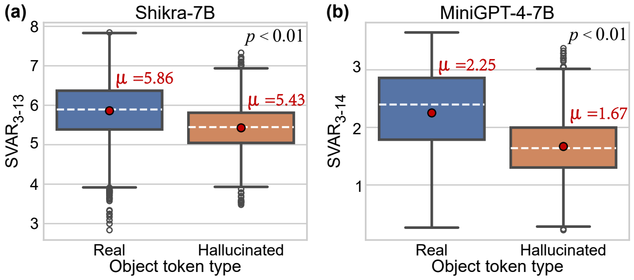


Figure 14. $\text{SVAR}_{3\text{-}13}$ and $\text{SVAR}_{3\text{-}14}$ score distributions across object token types for Shikra-7B (a) and MiniGPT-4-7B (b), respectively.

MHSA sublayers, respectively. We find the same two patterns in the middle layers analogous to those found in LLaVA-1.5-7B as described in Sec. 3.3, suggesting the model scale generalization of our findings. Fig. 15 presents qualitative comparisons of hallucination detection between the $\text{SVAR}_{5\text{-}18}$ metric and the internal confidence method, demonstrating the superiority of our metric.

MiniGPT-4-7B. Fig. 12 (a) and (b) depict the VAR score distribution and the prediction contributions from the MHSA sublayers, respectively. Similar to LLaVA-1.5-7B, the two patterns in the middle layers where the model exhibits continuous higher visual attention can be observed. Notably, we can see that MiniGPT-4-7B does not exhibit the same high attention as LLaVA-1.5 at the 0-th layer. In our experiments, layers 3-14 are selected as the range of the visual information enrichment stage. Fig. 14 (b) reports the $\text{SVAR}_{3\text{-}14}$ value distribution across the two token types, demonstrating a similar trend to LLaVA-1.5-7B as described in Sec. 3.4. These results suggest the model generalization of our findings. The qualitative results of hallucination detection are presented in Fig. 16.

Shikra-7B. As shown in Fig. 13 (a), Shikra continuously exhibits extremely high VAR scores across layers. To clearly analyze the VAR distribution, a seventh-order polynomial is used to fit the summed VAR values over all heads of each layer (depicted by a red curve). Compared to other layers, we can see that the middle layers exhibit relatively higher VAR scores, aligning with our observation from LLaVA-1.5-7B. Combined with the prediction contributions from MHSA sublayers in Fig. 13 (b), we can also identify two distinct patterns in the middle layers. Like MiniGPT-4-7B, Shikra-7B exhibits low visual attention at the 0-th layer. In our experiments, layers 3-13 are selected as the range of the visual information enrichment stage. Fig. 14 (a) presents the $\text{SVAR}_{3\text{-}13}$ value distribution across the two token types, demonstrating a similar trend to other LVLMs. The comparison results of hallucination detection, displayed in Fig. 17, show that our simple $\text{SVAR}_{3\text{-}13}$ metric performs comparably to the more complex baseline that projects the hidden states of all image tokens at all layers into the vocabulary space. Compared to other LVLMs, the decreased performance of the SVAR metric on Shikra-7B may be attributed to the extremely high VAR scores across nearly all layers, potentially reducing the sensitivity of our metric to attention pattern differences.

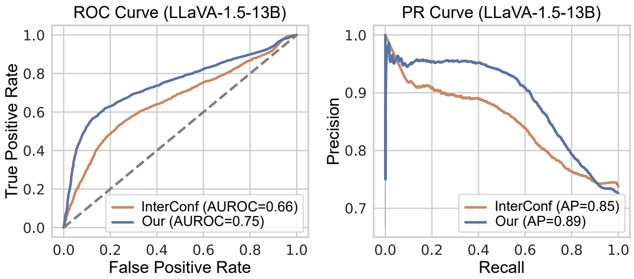


Figure 15. Object hallucinations detection curves for LLaVA-1.5-13B.

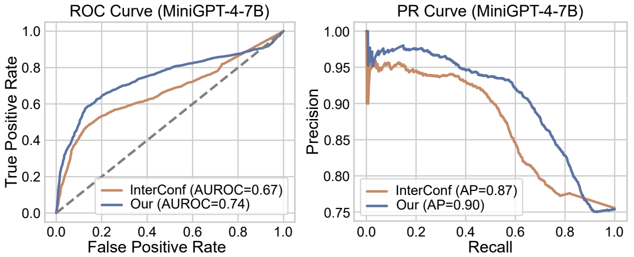


Figure 16. Object hallucinations detection curves for MiniGPT-4-7B.

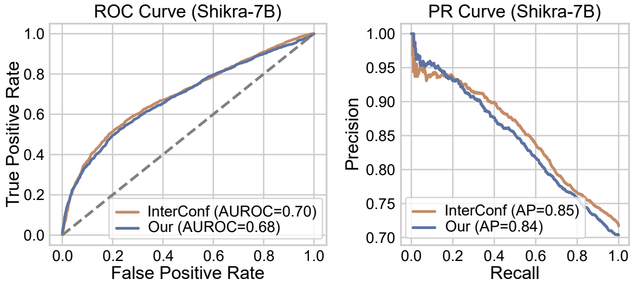


Figure 17. Object hallucinations detection curves for Shikra-7B.

LLaVA-1.5-7B	Real	Hallucinated
SVAR$_{5\text{-}18}$ score	1.70	1.25
t-statistic	32.44	
p-value	1.03E-213	
df	6,237	

Table 8. Results of one-tailed t-tests for LLaVA-1.5-7B. The null hypothesis is the mean SVAR$_{5\text{-}18}$ score of real object tokens is less than or equal to the mean SVAR$_{5\text{-}18}$ score of hallucinated ones.

LLaVA-1.5-13B	Real	Hallucinated
SVAR$_{5\text{-}18}$ score	1.50	1.06
t-statistic	32.24	
p-value	3.25E-211	
df	6,186	

Table 9. Results of one-tailed t-tests for LLaVA-1.5-13B. The null hypothesis is the mean SVAR$_{5\text{-}18}$ score of real object tokens is less than or equal to the mean SVAR$_{5\text{-}18}$ score of hallucinated ones.

Shikra-7B	Real	Hallucinated
SVAR$_{3\text{-}13}$ score	5.86	5.43
t-statistic	22.56	
p-value	1.50E-108	
df	6,055	

Table 10. Results of one-tailed t-tests for Shikra-7B. The null hypothesis is the mean SVAR$_{3\text{-}13}$ score of real object tokens is less than or equal to the mean SVAR$_{3\text{-}13}$ score of hallucinated ones.

MiniGPT-4-7B	Real	Hallucinated
SVAR$_{3\text{-}14}$ score	2.25	1.67
t-statistic	22.06	
p-value	4.21E-102	
df	3,978	

Table 11. Results of one-tailed t-tests for MiniGPT-4-7B. The null hypothesis is the mean SVAR$_{3\text{-}14}$ score of real object tokens is less than or equal to the mean SVAR$_{3\text{-}14}$ score of hallucinated ones.

C.2. Results of Statistical Tests

To assess the statistical significance of the SVAR score being higher for real object tokens than for hallucinated ones during visual information enrichment, we conduct a one-tailed t-test for each LVLM. We present the results in Tab. 8 for LLaVA-1.5-7B, Tab. 9 for LLaVA-1.5-13B, Tab. 10 for Shikra-7B, and Tab. 11 for MiniGPT-4-7B. Across all models, the results consistently indicate that significantly higher attention weights are assigned to image tokens when generating real object tokens, compared to hallucinated ones.

C.3. Numerical Results of α Sensitivity

Tab. 12 presents the sensitivity results of the balance factor α, used in our attention intervention method (Eq. (6)), on LLaVA-1.5-7B, LLaVA-1.5-13B, and MiniGPT-4-7B. In addition to modulating the trade-off between hallucination mitigation and description richness as discussed in Sec. 5, we find that LLaVA-1.5-7B and LLaVA-1.5-13B are more sensitive to changes in α compared to MiniGPT-4-7B. A possible reason for this increased sensitivity may be that LLaVA-1.5 uses substantially more image tokens than MiniGPT-4 (576 versus 32), potentially magnifying the impact of the parameter α.

C.4. Attention Heads Behavior Visualization

We exhibit more visualization examples of LLaVA-1.5-7B in Figs. 18 and 19 to validate that the heads interact with

α	LLaVA-1.5-7B			LLaVA-1.5-13B			MiniGPT-4-7B		
	$C_S \downarrow$	$C_I \downarrow$	F1↑	$C_S \downarrow$	$C_I \downarrow$	F1↑	$C_S \downarrow$	$C_I \downarrow$	F1↑
Greedy	53.0	15.6	76.7	49.8	14.6	78.2	31.8	12.0	71.1
0.3	41.8	12.2	77.9	44.4	12.5	78.1	28.0	10.0	70.5
0.4	41.6	11.5	78.1	44.6	13.2	77.5	27.2	10.8	71.2
0.5	25.0	6.7	76.1	25.8	8.8	77.3	22.4	8.6	70.8
0.6	1.0	0.9	49.3	6.4	3.3	57.7	20.6	8.6	69.5
0.7	1.6	2.0	36.8	2.4	26.3	40.0	14.6	5.9	67.4

Table 12. Numerical results of balance factor α sensitivity.

LLaVA-1.5-7B	Greedy	Beam	OPERA	VCD†	PAI†	**Ours**
CHAIR↓	7.7	9.1	7.3	8.6	4.9	**4.3**
Hal↓	35.4	39.8	31.5	39.5	24.4	**20.2**
Cog↓	4.3	4.8	2.9	4.5	1.6	**1.2**

Table 13. AMBER results on LLaVA-1.5-7B with *max new token* set to 512. † denotes using the greedy decoding strategy.

inconsistent objects in the image during visual information enrichment when generating hallucinated object tokens.

C.5. Comparison Results on AMBER Benchmark

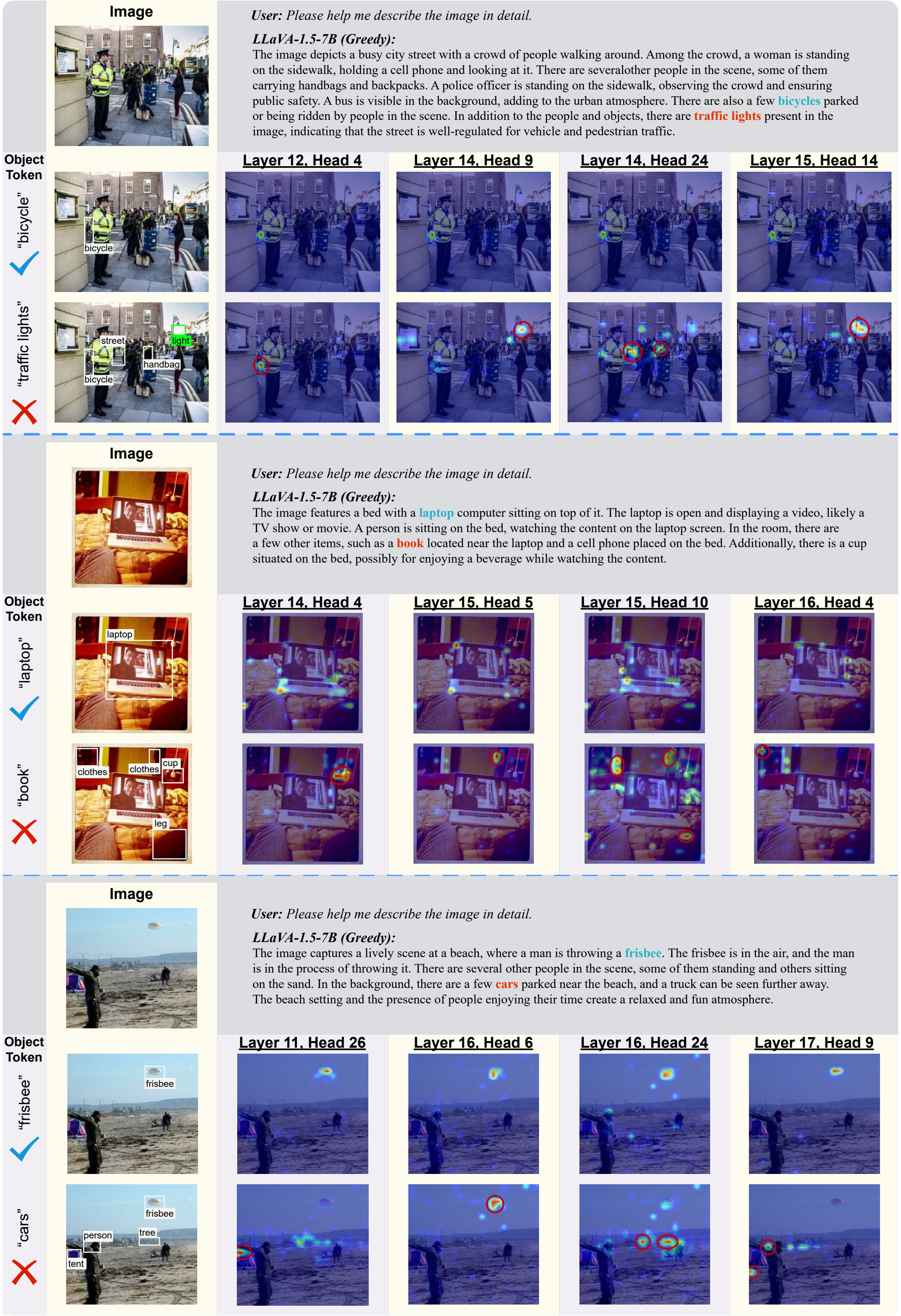
We further evaluate our approach on AMBER [41] benchmark, which contains 1,004 images for the generative task. The results presented in Tab. 13 demonstrate its superior performance.

C.6. Some Intuition Behind Layer Division

The layers of LLaVA-1.5-7B are divided into four ranges according to the patterns identified from the results of VAR score (Fig. 2 (a)) and logit lens (Fig. 2 (b) and Fig. 3). We find **Range 1** (layer 0-4): low-level image processing, VAR attention pattern differs in different models; **Range 2** (layer 5-18): visual information enrichment, accumulates the visual information exhibiting high VAR scores and low logit contribution; **Range 3** (layer 19-26): semantic refinement, interacts semantic information of image tokens with high VAR scores and reasons object token prediction with high logit contribution; **Range 4** (layer 27-31): grammar concern, guarantees the coherence and correctness of response with low VAR scores, in which the model tends to interpret image tokens as punctuation marks or conjunctions (Fig. 3). These patterns can be generalized to divide other LVLMs similarly. It is worth noting that the first and last divisions are not fixed at 5 layers but vary with models.

C.7. Qualitative Results of Hallucination Mitigation

We provide sample results from our hallucination mitigation method as described in Sec. 4.1, which corrects the attention distribution over image tokens, in Fig. 20, Fig. 21 and Fig. 22 for LLaVA-1.5-7B, MiniGPT-4-7B, and Shikra-7B, respectively.



User: *Please help me describe the image in detail.*

LLaVA-1.5-7B (Greedy):
The image depicts a busy city street with a crowd of people walking around. Among the crowd, a woman is standing on the sidewalk, holding a cell phone and looking at it. There are severalother people in the scene, some of them carrying handbags and backpacks. A police officer is standing on the sidewalk, observing the crowd and ensuring public safety. A bus is visible in the background, adding to the urban atmosphere. There are also a few **bicycles** parked or being ridden by people in the scene. In addition to the people and objects, there are **traffic lights** present in the image, indicating that the street is well-regulated for vehicle and pedestrian traffic.

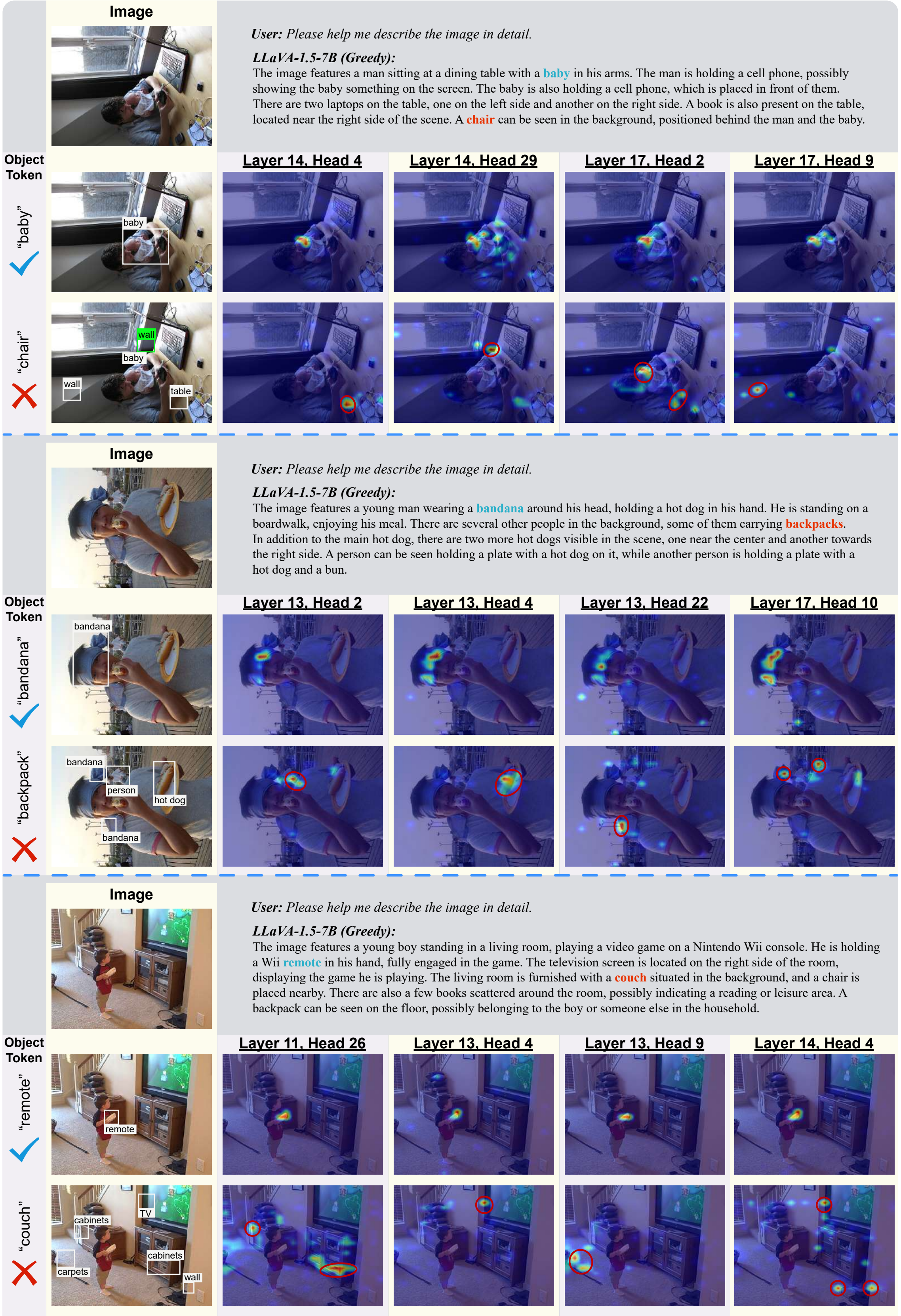
User: *Please help me describe the image in detail.*

LLaVA-1.5-7B (Greedy):
The image features a bed with a **laptop** computer sitting on top of it. The laptop is open and displaying a video, likely a TV show or movie. A person is sitting on the bed, watching the content on the laptop screen. In the room, there are a few other items, such as a **book** located near the laptop and a cell phone placed on the bed. Additionally, there is a cup situated on the bed, possibly for enjoying a beverage while watching the content.

User: *Please help me describe the image in detail.*

LLaVA-1.5-7B (Greedy):
The image captures a lively scene at a beach, where a man is throwing a **frisbee**. The frisbee is in the air, and the man is in the process of throwing it. There are several other people in the scene, some of them standing and others sitting on the sand. In the background, there are a few **cars** parked near the beach, and a truck can be seen further away. The beach setting and the presence of people enjoying their time create a relaxed and fun atmosphere.

Figure 18. Attention heads behavior in LLaVA-1.5-7B.



User: *Please help me describe the image in detail.*

LLaVA-1.5-7B (Greedy):
The image features a man sitting at a dining table with a **baby** in his arms. The man is holding a cell phone, possibly showing the baby something on the screen. The baby is also holding a cell phone, which is placed in front of them. There are two laptops on the table, one on the left side and another on the right side. A book is also present on the table, located near the right side of the scene. A **chair** can be seen in the background, positioned behind the man and the baby.

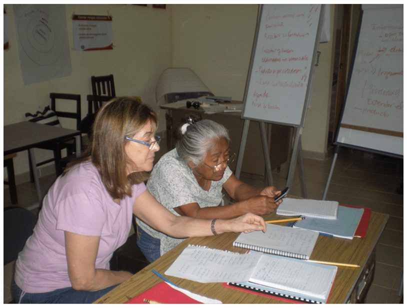
User: *Please help me describe the image in detail.*

LLaVA-1.5-7B (Greedy):
The image features a young man wearing a **bandana** around his head, holding a hot dog in his hand. He is standing on a boardwalk, enjoying his meal. There are several other people in the background, some of them carrying **backpacks**.
In addition to the main hot dog, there are two more hot dogs visible in the scene, one near the center and another towards the right side. A person can be seen holding a plate with a hot dog on it, while another person is holding a plate with a hot dog and a bun.

User: *Please help me describe the image in detail.*

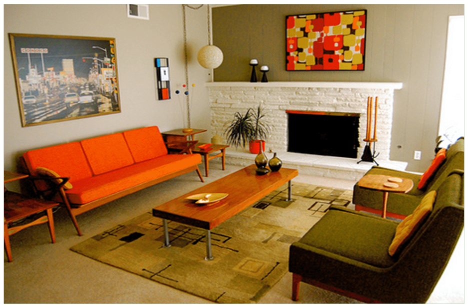
LLaVA-1.5-7B (Greedy):
The image features a young boy standing in a living room, playing a video game on a Nintendo Wii console. He is holding a Wii **remote** in his hand, fully engaged in the game. The television screen is located on the right side of the room, displaying the game he is playing. The living room is furnished with a **couch** situated in the background, and a chair is placed nearby. There are also a few books scattered around the room, possibly indicating a reading or leisure area. A backpack can be seen on the floor, possibly belonging to the boy or someone else in the household.

Figure 19. Attention heads behavior in LLaVA-1.5-7B.



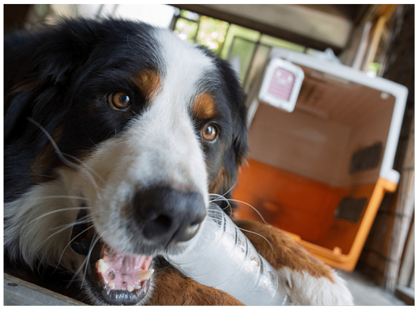
Greedy: The image features two **women** sitting at a dining table, engaged in a conversation. They are both wearing glasses and appear to be discussing something important. The table is filled with various items, including a **book**, a **cell phone**, and a notebook. There are several chairs around the **table**, with one on the left side and another on the right side. A **cup** can be seen placed on the table, and a **remote control** is also present. The room has a cozy atmosphere, with a **couch** in the background and a **TV** mounted on the wall.

Ours: In the image, two **women** are sitting at a table, with one woman looking at a **book** and the other woman looking at her cell phone. The table is covered with a red tablecloth, and there are several books on the **table**. The women are surrounded by a chair, a book, and a **cell phone**. The table is also equipped with a **laptop** and a notebook.



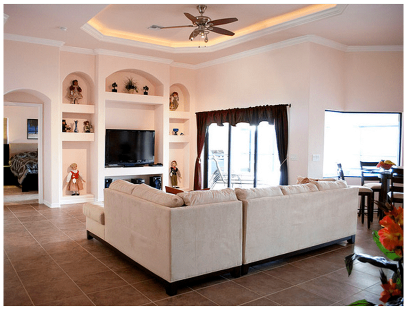
Greedy: The image depicts a spacious living room with a variety of furniture and decor. The room features a large **couch** situated in the middle, with a **chair** placed nearby. A **dining table** is also present in the room, accompanied by a few chairs. The living room is adorned with a **fireplace**, which serves as a focal point in the space. A potted plant is placed near the fireplace, adding a touch of greenery to the room. In addition to the main furniture, there are two **people** in the room, one standing near the left side and the other closer to the right side. A **car** can be seen in the background, possibly indicating that the living room is located in an apartment or a house with a garage.

Ours: The living room is decorated with a variety of furniture and artwork. There is a **couch**, a **chair**, and a **table** in the room. The couch is orange, and the chair is green. The room also features a **fireplace**, a potted plant, and a painting on the wall. The room is decorated with a variety of furniture and artwork, creating a cozy and inviting atmosphere.



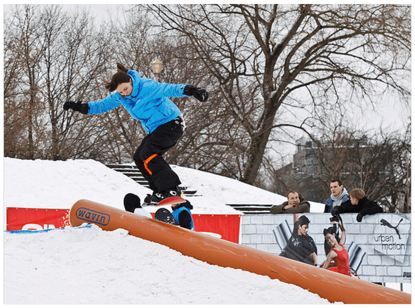
Greedy: The image features a large brown and white **dog** sitting on a table, with its mouth open and tongue hanging out. The dog appears to be enjoying its time and possibly panting. The **table** is located in a room with a window, providing natural light to the scene. In the background, there are two **people** present, one standing closer to the left side of the image and the other person standing further back on the right side. Additionally, there is a **chair** placed near the right side of the room.

Ours: The **dog** is a large, brown and white dog with a black nose. It is standing in front of a brown **box**, and its mouth is open. The dog appears to be a large breed, possibly a Saint Bernard. The dog is looking at the camera, possibly a picture of a dog.



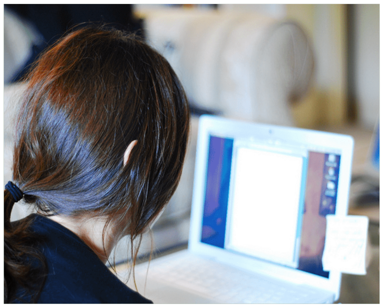
Greedy: The image depicts a spacious and well-furnished living room with a large **couch** and a **television**. The couch is positioned in the center of the room, with a **chair** placed nearby. The television is mounted on the wall, and there is a **remote control** placed on the couch. The living room is adorned with several decorative elements, including a **potted plant**, a **vase**, and a **clock**. The room also features a **dining table** with chairs, and a doorway leading to another room. The overall atmosphere of the living room is inviting and comfortable.

Ours: The image showcases a spacious living room with a large **couch** and a **television**. The couch is situated in the center of the room, and the television is placed on the left side of the room. The living room also features a **dining table** and **chairs**, with one chair visible in the scene. The room is well-decorated with a **potted plant** and a **vase**, adding a touch of greenery to the space. The room also has a large **window**, allowing natural light to fill the room.



Greedy: The image captures a **snowboarder** skillfully riding down a snow-covered ramp, performing a trick as they slide down the slope. The snowboarder is wearing a blue jacket and is in the middle of the action. There are several other **people** in the scene, some of them standing near the ramp, while others are scattered around the area. A **bench** can be seen in the background, and a **handbag** is placed on the ground near one of the people. The overall atmosphere of the scene is lively and energetic, with people enjoying the snowboarding activities.

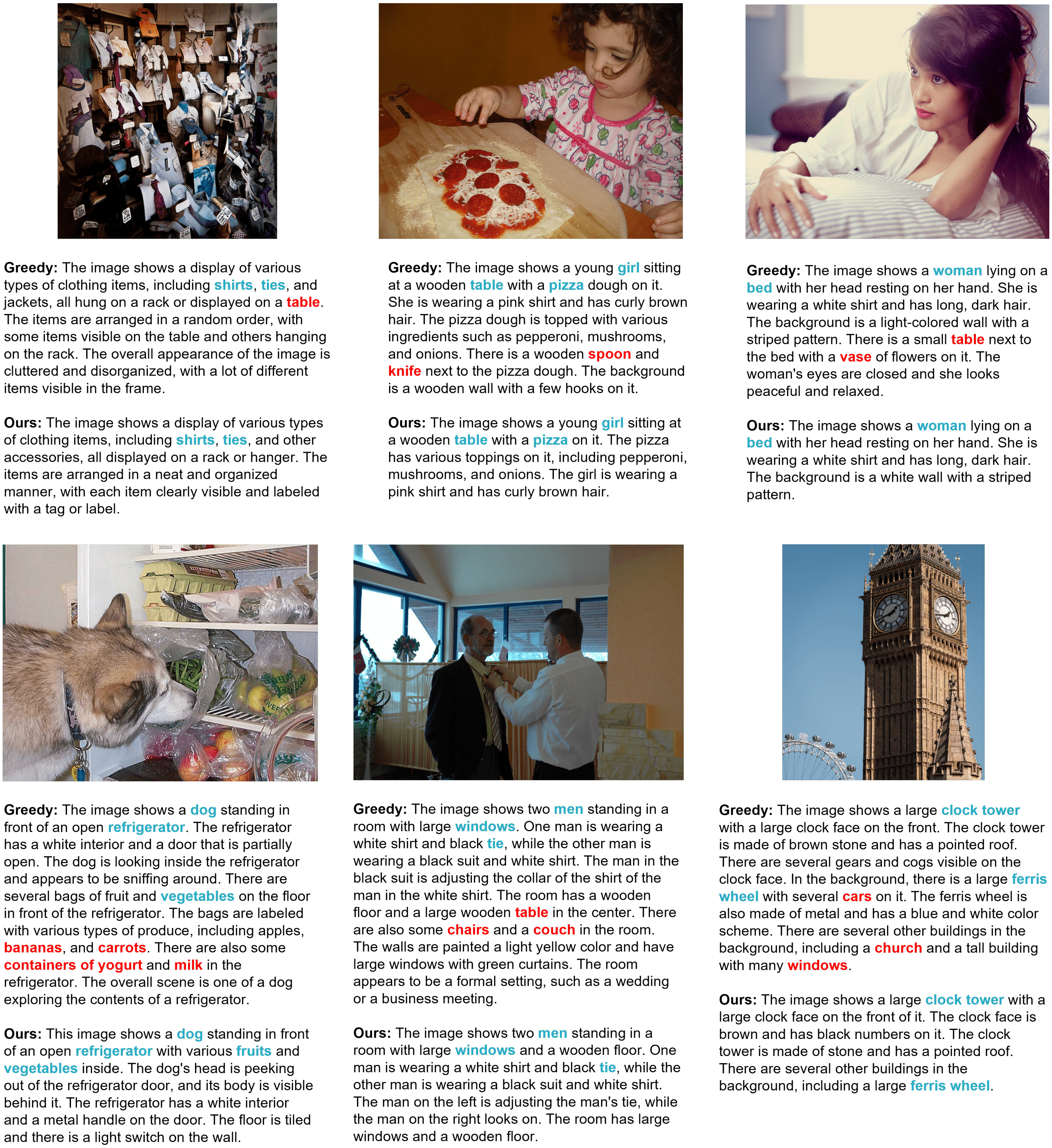
Ours: The image shows a **snowboarder** in a blue jacket riding a snowboard down a ramp. The snowboarder is wearing a blue jacket and is performing a trick on the ramp. There are several **people** watching the snowboarder, including a man in a black jacket and a woman in a red jacket. The snowboarder is in the middle of the ramp, and the spectators are spread out around the scene, with some standing closer to the ramp and others further away.



Greedy: The image features a **woman** sitting in front of a **laptop** computer, which is placed on a desk. She is focused on the screen, likely working or browsing the internet. The woman is wearing a black shirt and has her hair pulled back in a ponytail. In the background, there is a **couch** and a **chair**, providing a comfortable seating area for the woman. Additionally, there are two **bottles** placed on the **desk**, one near the laptop and the other further away. A **cell phone** can also be seen on the desk, possibly for communication or reference purposes.

Ours: The image shows a **woman** sitting in front of a **laptop** computer. She is wearing a black shirt and has her hair pulled back in a ponytail. The laptop screen is on, and she is looking at it. The laptop is white and is placed on a **table**.

Figure 20. Qualitative results of hallucination mitigation on LLaVA-1.5-7B. The real and hallucinated object words are marked in blue and red, respectively. Our prompt is “Please help me describe the image in detail.”.



Greedy: The image shows a display of various types of clothing items, including **shirts**, **ties**, and jackets, all hung on a rack or displayed on a **table**. The items are arranged in a random order, with some items visible on the table and others hanging on the rack. The overall appearance of the image is cluttered and disorganized, with a lot of different items visible in the frame.

Ours: The image shows a display of various types of clothing items, including **shirts**, **ties**, and other accessories, all displayed on a rack or hanger. The items are arranged in a neat and organized manner, with each item clearly visible and labeled with a tag or label.

Greedy: The image shows a young **girl** sitting at a wooden **table** with a **pizza** dough on it. She is wearing a pink shirt and has curly brown hair. The pizza dough is topped with various ingredients such as pepperoni, mushrooms, and onions. There is a wooden **spoon** and **knife** next to the pizza dough. The background is a wooden wall with a few hooks on it.

Ours: The image shows a young **girl** sitting at a wooden **table** with a **pizza** on it. The pizza has various toppings on it, including pepperoni, mushrooms, and onions. The girl is wearing a pink shirt and has curly brown hair.

Greedy: The image shows a **woman** lying on a **bed** with her head resting on her hand. She is wearing a white shirt and has long, dark hair. The background is a light-colored wall with a striped pattern. There is a small **table** next to the bed with a **vase** of flowers on it. The woman's eyes are closed and she looks peaceful and relaxed.

Ours: The image shows a **woman** lying on a **bed** with her head resting on her hand. She is wearing a white shirt and has long, dark hair. The background is a white wall with a striped pattern.

Greedy: The image shows a **dog** standing in front of an open **refrigerator**. The refrigerator has a white interior and a door that is partially open. The dog is looking inside the refrigerator and appears to be sniffing around. There are several bags of fruit and **vegetables** on the floor in front of the refrigerator. The bags are labeled with various types of produce, including apples, **bananas**, and **carrots**. There are also some **containers of yogurt** and **milk** in the refrigerator. The overall scene is one of a dog exploring the contents of a refrigerator.

Ours: This image shows a **dog** standing in front of an open **refrigerator** with various **fruits** and **vegetables** inside. The dog's head is peeking out of the refrigerator door, and its body is visible behind it. The refrigerator has a white interior and a metal handle on the door. The floor is tiled and there is a light switch on the wall.

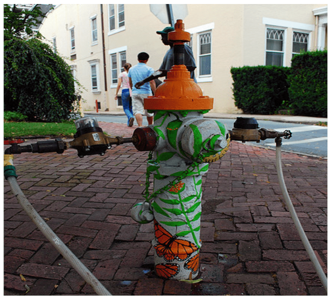
Greedy: The image shows two **men** standing in a room with large **windows**. One man is wearing a white shirt and black **tie**, while the other man is wearing a black suit and white shirt. The man in the black suit is adjusting the collar of the shirt of the man in the white shirt. The room has a wooden floor and a large wooden **table** in the center. There are also some **chairs** and a **couch** in the room. The walls are painted a light yellow color and have large windows with green curtains. The room appears to be a formal setting, such as a wedding or a business meeting.

Ours: The image shows two **men** standing in a room with large **windows** and a wooden floor. One man is wearing a white shirt and black **tie**, while the other man is wearing a black suit and white shirt. The man on the left is adjusting the man's tie, while the man on the right looks on. The room has large windows and a wooden floor.

Greedy: The image shows a large **clock tower** with a large clock face on the front. The clock tower is made of brown stone and has a pointed roof. There are several gears and cogs visible on the clock face. In the background, there is a large **ferris wheel** with several **cars** on it. The ferris wheel is also made of metal and has a blue and white color scheme. There are several other buildings in the background, including a **church** and a tall building with many **windows**.

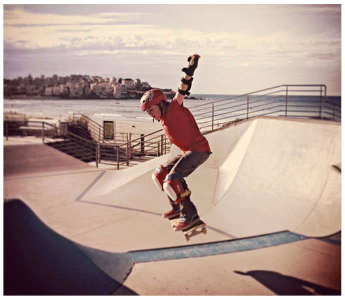
Ours: The image shows a large **clock tower** with a large clock face on the front of it. The clock face is brown and has black numbers on it. The clock tower is made of stone and has a pointed roof. There are several other buildings in the background, including a large **ferris wheel**.

Figure 21. Qualitative results of hallucination mitigation on MiniGPT-4-7B. The real and hallucinated object words are marked in blue and red, respectively. Our prompt is “Please help me describe the image in detail.”.



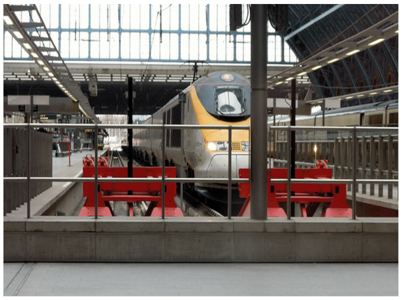
Greedy: The image features a fire hydrant that has been painted in a vibrant pattern, including a mix of colors such as orange, green, and white. The fire hydrant is located on a brick sidewalk, which appears to be a part of a street. There are two people in the scene, one standing closer to the fire hydrant while the other is further away. The person standing near the hydrant is carrying a handbag. Additionally, there is a car parked on the street, situated in the background.

Ours: The image features a fire hydrant that has been painted in a bright, colorful pattern. The hydrant is located on a brick sidewalk, with people walking nearby. There are two people walking in the background, likely enjoying the view of the colorful fire hydrant. The fire hydrant is connected to a water hose that is positioned close to the right side of the image.



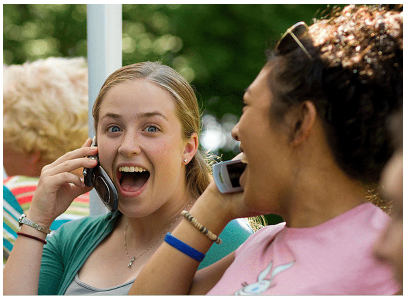
Greedy: The image showcases a young boy wearing a red shirt and helmet, skateboarding on a ramp near the ocean. He is in the middle of performing a skateboard trick, possibly jumping up in the air. In the background, there are several cars and a bus parked close to the ramp, possibly belonging to the visitors or spectators of the skate park. The ocean can be seen beyond the ramp, creating a picturesque setting for the skateboarding activities.

Ours: The image showcases a young boy riding a skateboard and jumping up in the air while wearing a helmet at a skate park or a ramp. The boy is wearing a red shirt and is fully geared up with protective gear, indicating that he might be practicing skateboarding tricks. In the background, there is a beach visible, adding a relaxing and enjoyable atmosphere to the scene. The young boy is likely having fun while practicing his skateboarding skills on a sunny day.



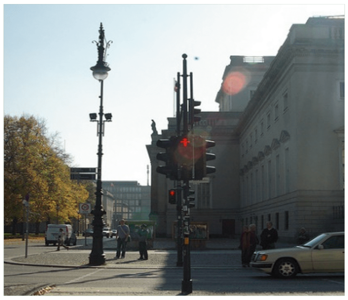
Greedy: The image features a modern train station with a sleek-looking train positioned on the tracks. The train is surrounded by several benches, with four chairs located in front of it, providing seating for passengers. There are also three other benches situated in various parts of the station. The station is bustling with activity, as a group of people can be seen standing and walking around the area. Some people are close to the train, while others are further away, either waiting for the train or simply enjoying their time at the station. Overall, the scene captures the lively atmosphere of a busy train station.

Ours: The image showcases a modern train station with a sleek-looking black and yellow train sitting in the background. The train is parked on the tracks, and it appears to be stationary inside the building. The train occupies a significant portion of the scene, taking up a majority of the image.



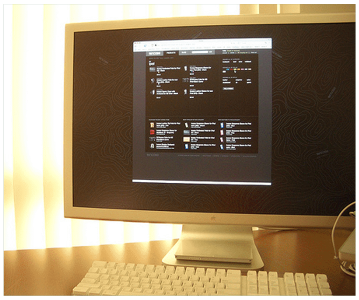
Greedy: The image features two young women sitting together in an outdoor setting, each talking on their cell phones. One woman is on her phone, while the other is using a cell phone that is not visible in the scene. They seem to be engaged in separate conversations or enjoying their time together. Additionally, there are a few other people in the scene, but they are not the main focus of the image. A chair can be seen in the background, and it appears that the women are seated on it while they chat on their phones.

Ours: The image shows two young women sitting in a park, each holding a cell phone to their ears. One of them is talking on her phone while the other is holding a phone. One of them is making a funny face as they are engaged in conversation. In the scene, there is an older woman in the background who is not visible. The young women are sitting together, laughing and talking on their cell phones. The image focuses on the interaction between the two girls, each with their own cell phone.



Greedy: The scene depicts a busy city intersection with a red traffic light near a large building. Several people are walking on the street, some carrying handbags. There are five pedestrians in total, with some walking closer to the traffic light while others are further away. Two cars are visible in the scene, one positioned near the traffic light and another one closer to the right side of the image. A bus can also be seen in the background, likely providing public transportation for people in the area. A streetlight is present in the scene, illuminating the area for safe navigation during nighttime or low-light conditions. Additionally, a bicycle is parked near the traffic light, hinting at the various modes of transportation used in the city.

Ours: The scene features a city street with a red traffic light showing a car stopped at a street corner. There are several people walking and standing near the traffic light, some of them are waiting to cross the street. The traffic light is located at an intersection in a city setting. In the image, there are multiple traffic lights placed at different positions along the road. One of the traffic lights is red, while the other traffic lights are placed at various locations along the street. The pedestrians are waiting to cross the street, while the car is stopped at the intersection.



Greedy: The image features a large computer monitor sitting on a wooden desk, with a keyboard placed in front of it. The monitor displays a black box, which seems to be a home screen for a computer. There is also a white mouse located on the right side of the keyboard. The desk is well-organized, and a book can be seen on the left side of the keyboard. Additionally, there are two cups placed on the desk, one near the center and another towards the right side. The scene suggests a workspace with a computer setup for daily use.

Ours: The image features a large computer monitor displaying a wide variety of items on the screen. There is a keyboard on the desk in front of the monitor, which indicates that the user can type and navigate through the computer. The monitor takes up most of the space on the desk, while the keyboard occupies a significant portion of the desk area.

Figure 22. Qualitative results of hallucination mitigation on Shikra-7B. The real and hallucinated object words are marked in blue and red, respectively. Our prompt is "Please help me describe the image in detail.".