

Prefill-Time Intervention for Mitigating Hallucination in Large Vision-Language Models

Supplementary Material

A. Details of Internal Interpretability Analysis.

In this section, we provide the mathematical formulation and calculation steps of the change rate of attention for the internal interpretability analysis presented in Section 5.2.

Specifically, we randomly sample 300 images from the MSCOCO [24] dataset. For each image, we denote $\mathbf{A}_{l,h}^t \in \mathbb{R}^N$ as the last token attention distribution of the h -th head in the l -th layer at generation step t over the sequence of length N , \mathcal{I}_{img} represent the set of indices corresponding to visual tokens. Then, we quantitatively measure the rate of change in both 1) Global Visual Attention and 2) Local Object-Centric Attention before and after our intervention.

A.1. Global Visual Attention Dynamics

To quantify how the model’s reliance on visual information evolves throughout the generation process, we employ a stage-wise attention analysis. This allows us to monitor the “attention decay” phenomenon [20, 30, 37] and verify whether our intervention effectively counteracts this trend.

Visual Attention Proportion. First, for any given generation step t , we define the global visual attention proportion P_{img}^t as follows:

$$P_{\text{img}}^t = \frac{1}{L \cdot H} \sum_{l=1}^L \sum_{h=1}^H \frac{\sum_{v \in \mathcal{I}_{\text{img}}} \mathbf{A}_{l,h}^t[v]}{\sum_{n \in N} \mathbf{A}_{l,h}^t[n] + \epsilon} \quad (9)$$

which represents the average probability mass allocated to visual tokens across all layers L and heads H . Here, ϵ is a small constant for numerical stability.

Temporal Alignment via Staged Sampling. Since the total length of generated tokens N varies between the original (before) and intervened (after) models, we normalize the generation process into discrete progression stages to ensure a fair alignment. Specifically, we divide the generation into K discrete stages. For the k -th stage, we sample the proportion of visual attention at step t_k as follows:

$$t_k = \lfloor \frac{k}{K} \cdot N \rfloor, \quad k \in \{0, 1, \dots, K\} \quad (10)$$

which effectively captures the behavior of the model from the onset (0%) to the completion (100%) of the response.

Global Relative Change Rate. To evaluate the impact of our intervention, we calculate the relative change in visual attention at each aligned stage. Let $P_{\text{img}}^{t_k}$ and $\hat{P}_{\text{img}}^{t_k}$ denote the visual attention proportions for the vanilla (before) and intervened (after) models at stage k , respectively. The stage-

wise relative change rate ΔR^k is defined as follows:

$$\Delta R^k = \frac{\hat{P}_{\text{img}}^{t_k} - P_{\text{img}}^{t_k}}{|P_{\text{img}}^{t_k}| + \epsilon} \times 100\%. \quad (11)$$

Here, a positive ΔR^k indicates that the intervention effectively enhances the model’s visual grounding capabilities relative to the Vanilla. Crucially, by analyzing the trend of $\Delta R = \{\Delta R^0, \Delta R^1, \dots, \Delta R^K\}$, we can verify whether the intervention specifically counteracts the attention decay that typically occurs in the later stages of long-form generation.

A.2. Local Object-Centric Attention Shift

Beyond preserving the global magnitude of visual signals, we further verify whether the enhanced visual attention is meaningfully focused on the relevant object regions rather than the background.

Visual Attention Re-normalization. We first disentangle the spatial distribution from the total visual attention weight. For a specific layer l and head h at step t , we compute the conditional probability distribution $\tilde{\mathbf{A}}_{l,h}^t$ over the visual tokens:

$$\tilde{\mathbf{A}}_{l,h}^t[v] = \frac{\mathbf{A}_{l,h}^t[v]}{\sum_{j \in \mathcal{I}_{\text{img}}} \mathbf{A}_{l,h}^t[j]}, \quad \forall v \in \mathcal{I}_{\text{img}} \quad (12)$$

This re-normalization reflects the spatial focus of the head, given that it is attending to the image.

Object-Centric Attention Score. Let $M \in \{0, 1\}^{|\mathcal{I}_{\text{img}}|}$ be the binary mask of the ground-truth object, where $M[v] = 1$ indicates token v falls within the object region. We calculate the aggregated attention mass allocated to the object as:

$$S_{\text{obj}}^t(l, h) = \sum_{v \in \mathcal{I}_{\text{img}}} \tilde{\mathbf{A}}_{l,h}^t[v] \cdot M[v] \quad (13)$$

This metric (S_{obj}) provides a granular view of the model’s spatial grounding capability at the head level.

Local Targeted Shift Finally, to visualize the layer-wise and head-wise impact of our intervention, we compute the absolute shift in object-centric attention: Let $S_{\text{obj}}^t(l, h)$ and $\hat{S}_{\text{obj}}^t(l, h)$ denote the proportions of object attention for the vanilla (before) and intervened (after) models at step t , respectively. The object-centric attention shift $\Delta S_{\text{obj}}^t(l, h)$ is defined as follows:

$$\Delta S_{\text{obj}}^t(l, h) = \hat{S}_{\text{obj}}^t(l, h) - S_{\text{obj}}^t(l, h). \quad (14)$$

Table 7. More results on the random and popular splits of the POPE benchmark.

| Method | LLAVA-1.5 | | | | Qwen-VL-Chat | | | | DeepSeek-VL-Chat | | | |
|-------------------|----------------|---------------|----------------|---------------|----------------|---------------|----------------|---------------|------------------|---------------|----------------|---------------|
| | Random | | Popular | | Random | | Popular | | Random | | Popular | |
| | Acc \uparrow | F1 \uparrow | Acc \uparrow | F1 \uparrow | Acc \uparrow | F1 \uparrow | Acc \uparrow | F1 \uparrow | Acc \uparrow | F1 \uparrow | Acc \uparrow | F1 \uparrow |
| Vanilla | 83.74 | 84.54 | 80.50 | 81.49 | 86.52 | 85.59 | 84.30 | 83.39 | 84.81 | 84.18 | 84.13 | 83.24 |
| PAI [30] | 85.73 | 86.32 | 82.53 | 83.40 | 87.69 | 86.94 | <u>85.40</u> | <u>84.40</u> | 86.83 | 86.29 | <u>85.86</u> | <u>84.93</u> |
| VCD [17] | 85.36 | <u>86.00</u> | <u>82.60</u> | <u>83.41</u> | 86.93 | 85.46 | 85.17 | 83.68 | 86.70 | 85.97 | 85.80 | 84.56 |
| VTI [29] | 83.40 | 84.14 | 81.50 | 82.33 | 85.77 | 84.68 | 83.86 | 82.59 | 86.15 | 86.01 | 84.83 | 84.48 |
| VISTA [23] | 84.74 | 85.78 | 79.97 | 81.55 | 85.91 | 84.66 | 83.90 | 82.25 | <u>86.70</u> | 85.54 | 86.13 | 84.58 |
| PTI (ours) | <u>85.43</u> | 85.72 | 83.80 | 84.09 | <u>87.53</u> | <u>86.54</u> | 86.17 | 84.89 | 86.22 | <u>86.25</u> | 85.43 | 85.16 |

Table 8. Evaluation of MME Benchmark. ‘‘Exist.’’: Existence. ‘‘Posit.’’: Position. The maximum new token is set to 32.

| Method | LLAVA-1.5 | | | | | Qwen-VL-Chat | | | | | DeepSeek-VL-Chat | | | | |
|-------------------|-----------|-------|--------|-------|--------------|--------------|-------|--------|-------|--------------|------------------|-------|--------|-------|--------------|
| | Exist. | Count | Posit. | Color | Total | Exist. | Count | Posit. | Color | Total | Exist. | Count | Posit. | Color | Total |
| Vanilla | 180.0 | 143.3 | 133.3 | 155.0 | 611.6 | 180.0 | 115.0 | 128.3 | 175.0 | 598.3 | 190.0 | 153.3 | 133.3 | 175.0 | 651.6 |
| PAI [30] | 190.0 | 148.3 | 126.6 | 160.0 | 625.0 | 180.0 | 120.0 | 130.0 | 175.0 | 605.0 | 190.0 | 158.3 | 133.3 | 175.0 | 656.6 |
| VTI [29] | 185.0 | 153.3 | 130.0 | 165.0 | <u>633.3</u> | 185.0 | 145.0 | 116.6 | 180.0 | <u>626.6</u> | 195.0 | 158.3 | 133.3 | 175.0 | <u>661.6</u> |
| VISTA [23] | 195.0 | 138.3 | 121.6 | 160.0 | 615.0 | 175.0 | 125.0 | 141.6 | 170.0 | 611.6 | 185.0 | 158.3 | 128.3 | 175.0 | 646.6 |
| PTI (ours) | 195.0 | 163.3 | 128.3 | 165.0 | 651.6 | 185.0 | 140.0 | 148.3 | 165.0 | 638.3 | 195.0 | 163.3 | 138.3 | 175.0 | 671.6 |

A positive $\Delta S_{\text{obj}}(l, h)$ signifies that the h -th head in the l -th layer has successfully redistributed probability mass from the background to the object region. We visualize these shifts as heatmaps (see Figure 5 Right) to identify which specific components of the Transformer are responsible for the improved object grounding.

B. Additional Experiments of PTI.

B.1. More Experimental Results of POPE.

Table 7 presents the detailed performance on the *Random* and *Popular* splits of the POPE benchmark, supplementing the Adversarial results discussed in Section 4.3. Consistent with the performance in Table 2, PTI demonstrates robust generalization capabilities across varying difficulty levels. Specifically, in the *Popular* split, which challenges models with high-frequency objects prone to statistical language priors, PTI achieves superior performance, surpassing the Vanilla baseline by substantial margins (e.g., +2.60% F1 on LLaVA-1.5 and +1.50% F1 on Qwen-VL-Chat) and outperforming competitive baselines like PAI [30] and VCD [17]. In the *Random* split, where performance is generally saturated due to lower difficulty, PTI maintains competitive results, consistently ranking within the top-two across all evaluated LLMs. Notably, PTI consistently outperforms both VTI [29] and VISTA [23], offering a significantly more effective steering paradigm.

B.2. Detailed Experimental Results of MME.

Table 8 details the complete performance on the MME benchmark for evaluating object-level and attribute-level hallucination. Notably, PTI exhibits substantial gains in the

most challenging fine-grained perception tasks, specifically ‘‘Count’’ and ‘‘Position’’ where standard models frequently struggle. While the competitive VTI method shows strength in specific attributes like color, PTI maintains a superior balance across all dimensions. This consistent superiority in spatially sensitive and quantitative metrics validates that our PTI effectively sharpens the model’s initial visual grounding, enabling it to better resolve distinct objects and their spatial relationships before the decoding phase begins.

B.3. Inference Efficiency Analysis.

As shown in Table 9, we evaluate the inference efficiency of our method against competing approaches in terms of both latency and throughput. Existing methods often introduce significant computational overhead. Methods apply case-specific operations, such as VCD [17] and PAI [30], introduce significant overhead due to contrastive computations, nearly doubling the latency (e.g., $\times 1.99$ for VCD on LLaVA-1.5). Consequently, their throughput suffers a substantial degradation, dropping to approximately $0.50\times$ of the performance of vanilla. Additionally, DTI methods like VTI [29], which apply steering vectors across multiple steps, also incur notable costs (e.g., $\times 1.07$, $\times 1.06$, and $\times 1.08$, respectively), resulting in a throughput decline lower than $\times 0.95$. In contrast, PTI eliminates the need for sample-specific operations or multi-step interventions by modifying the initial KV cache only once. As a result, it incurs negligible latency overhead (lower than $\times 1.02$ across all models), and sustains near-lossless throughput (over $\times 0.98$ across all models). This consistent superiority highlights that PTI is a highly efficient, plug-and-play solution suitable for delay-sensitive real-world applications.

Table 9. Measure of Latency (ms/token) and Throughput (tokens/s) on CHAIR benchmark. All results use the Nucleus Sampling decoding strategy on a NVIDIA 4090 GPU.

| Method | LLaVA-1.5 | | Qwen-VL-Chat | | DeepSeek-VL-Chat | |
|-------------------|---|---|---|---|---|---|
| | Latency ↓ | Throughput ↑ | Latency ↓ | Throughput ↑ | Latency ↓ | Throughput ↑ |
| Vanilla | 19.52 ($\times 1.00$) | 51.22 ($\times 1.00$) | 20.55 ($\times 1.00$) | 48.66 ($\times 1.00$) | 18.29 ($\times 1.00$) | 54.67 ($\times 1.00$) |
| VCD [17] | 38.92 ($\times 1.99$) | 25.69 ($\times 0.50$) | 41.08 ($\times 1.99$) | 24.34 ($\times 0.50$) | 36.38 ($\times 1.98$) | 27.49 ($\times 0.50$) |
| PAI [30] | 37.62 ($\times 1.93$) | 26.58 ($\times 0.52$) | 42.07 ($\times 2.04$) | 23.77 ($\times 0.48$) | 35.16 ($\times 1.92$) | 28.44 ($\times 0.52$) |
| VISTA [23] | 26.20 ($\times 1.34$) | 38.17 ($\times 0.74$) | 33.81 ($\times 1.64$) | 29.57 ($\times 0.60$) | 26.79 ($\times 1.46$) | 37.33 ($\times 0.68$) |
| VTI [29] | 20.84 ($\times 1.07$) | 47.97 ($\times 0.93$) | 21.86 ($\times 1.06$) | 45.75 ($\times 0.94$) | 19.81 ($\times 1.08$) | 50.47 ($\times 0.92$) |
| PTI (ours) | 19.58 ($\times 1.00$) | 51.06 ($\times 0.99$) | 20.88 ($\times 1.01$) | 47.90 ($\times 0.98$) | 18.56 ($\times 1.02$) | 53.89 ($\times 0.99$) |

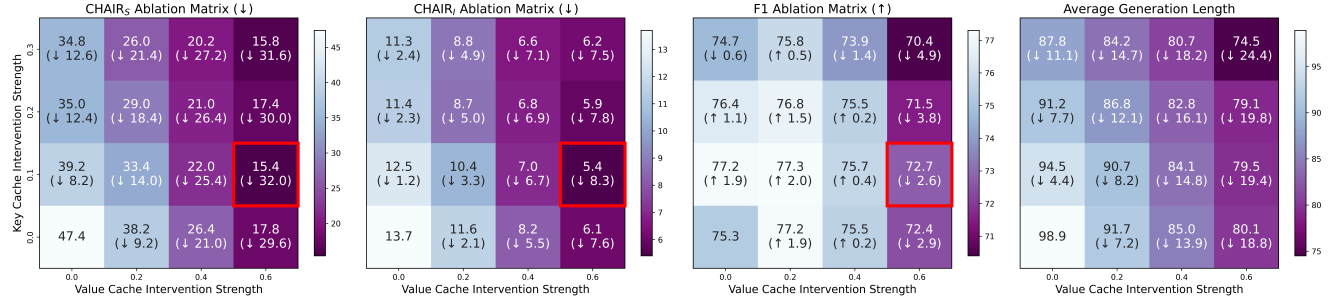


Figure 6. Ablation matrices for multi-modal KV cache intervention strength on LLaVA-1.5 with greedy decoding strategy. Brighter colors indicate better performance, while red boxes highlight the parameter combinations used in Table 1.

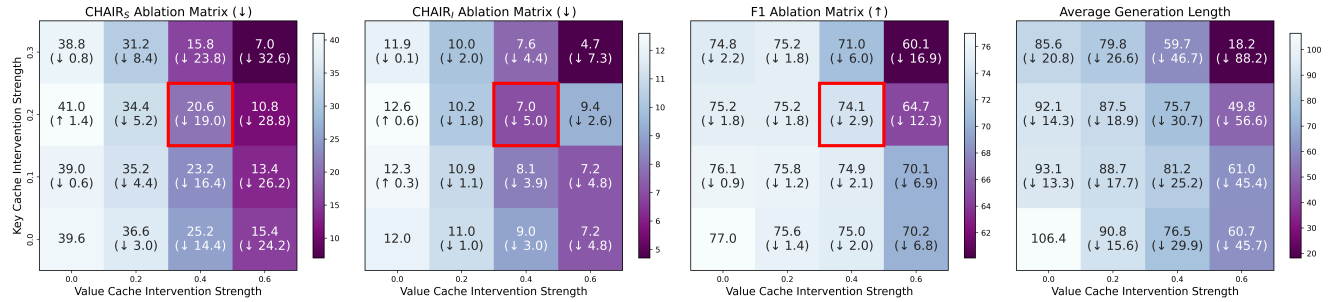


Figure 7. Ablation matrices for multi-modal KV cache intervention strength on Qwen-VL-Chat with greedy decoding strategy.

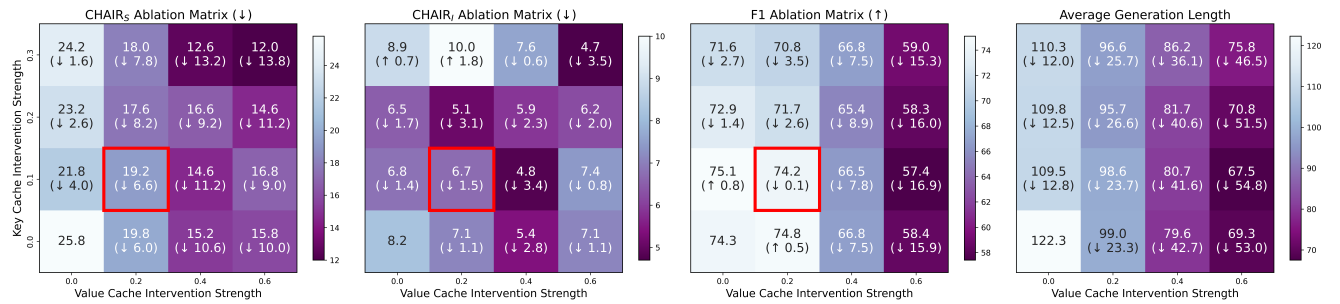


Figure 8. Ablation matrices for multi-modal KV cache intervention strength on DeepSeek-VL-Chat with greedy decoding strategy.

B.4. Ablation Experiments of Hyperparameters.

To investigate the sensitivity of PTI to intervention intensity, we present the ablation matrices for key and value intervention strengths on CHAIR [33] benchmark across LLaVA-1.5 (Figure 6), Qwen-VL-Chat (Figure 7), and

DeepSeek-VL-Chat (Figure 8). As mentioned in Section 4.2, we unify the visual and textual coefficients by setting $\lambda_{k, \text{img}} = \lambda_{k, \text{txt}}$ and $\lambda_{v, \text{img}} = \lambda_{v, \text{txt}}$ to reduce the hyperparameter search space. A consistent trend is observable across these architecture-distinct models: the value cache intervention exerts a more dominant influence on hal-

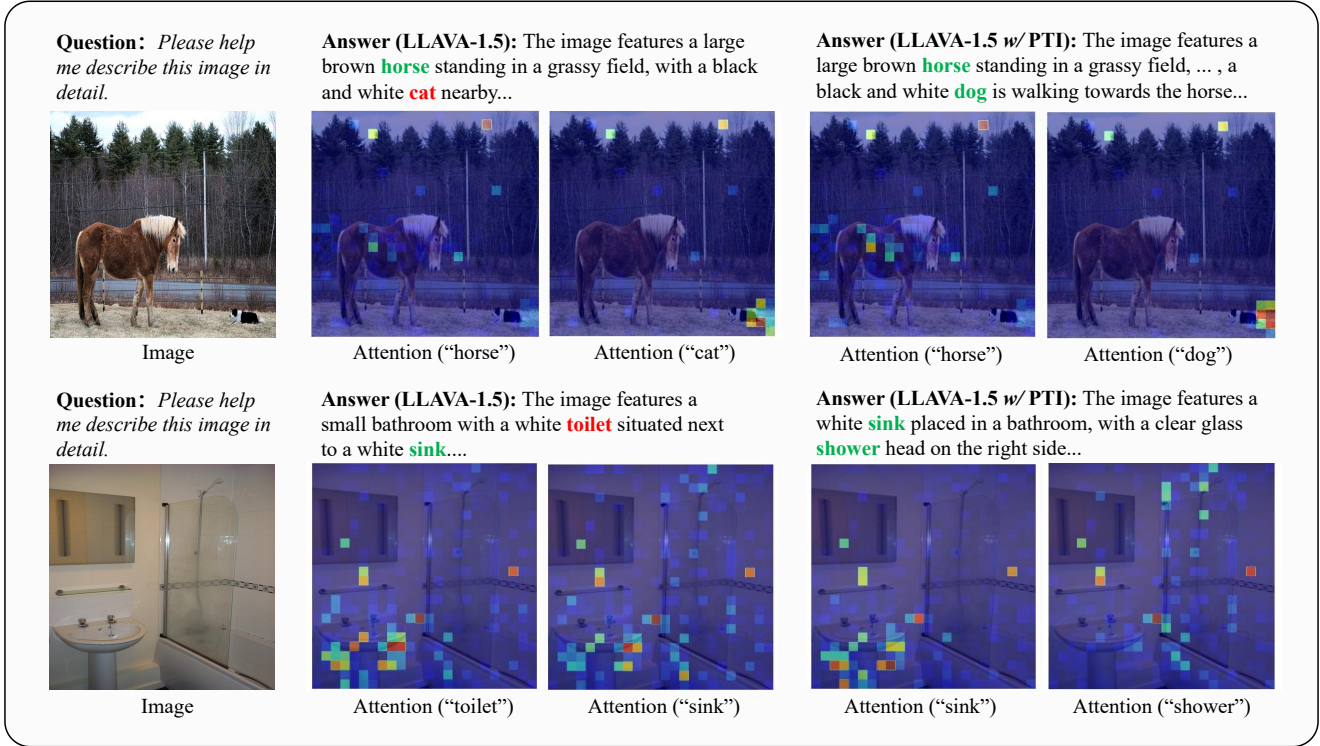


Figure 9. Visual analysis of cross-modal attention maps on LLaVA-1.5. For each sample, the hallucinated content is highlighted in **red**, while the correct content is highlighted in **green**. The attention map of the target token represents the magnitude of attention weights assigned to image tokens, averaged across all layers and heads.

lucination mitigation, as evidenced by the substantial reduction in $CHAIR_S/CHAIR_I$ metrics corresponding to increases in value intervention strength. However, the matrices also reveal a critical sensitivity to excessive steering; simultaneously maximizing both key and value intervention strengths precipitates a noticeable degradation in F1 scores and a sharp decline in average generation length, indicating a compromise in the model’s generation quality. Consequently, the configurations highlighted in red boxes represent an optimal equilibrium, effectively minimizing hallucination rates via robust visual grounding while preserving the fidelity and completeness of the textual response.

C. Additional Case Studies

C.1. Attention Map Visualization.

To intuitively understand how PTI mitigates hallucinations, we visualize the cross-modal attention maps of the generated tokens with respect to the visual features. As illustrated in Figure 9, PTI substantially enhances the object-centric attention and robust visual recognition of LLaVA-1.5. Taking the first example as prominent cases, the vanilla model suffers from severe perceptual misalignment and erroneously hallucinates a “cat”. In contrast, PTI effectively corrects

this distribution. On one hand, PTI intensifies the attentional weights on the correctly identified dominant object (“horse”), successfully rectifying the model’s gaze. On the other hand, PTI ensures a more robust and stable visual representation, enabling the model to identify the “dog” correctly. This dual effect—strengthening valid signals while correcting misaligned features—confirms that PTI operates by enforcing precise object-centric attention, thereby eliminating object-level hallucinations at their source.

C.2. Qualitative Examples.

We further extend our qualitative evaluation across all three distinct architectures to demonstrate PTI’s effectiveness in reducing hallucination. Figures 10-12 present comparative examples between vanilla models, DTI methods (*i.e.*, VTI [29] and VISTA [23]), and our PTI for LLaVA-1.5, Qwen-VL-Chat, and DeepSeek-VL-Chat, respectively. As evident across these scenarios, while vanilla models and DTI methods frequently suffer from severe object hallucinations and context misinterpretation, PTI effectively suppresses the generation of non-existent entities and erroneous attributes. These qualitative examples demonstrate the superior ability of PTI to mitigate hallucination in large vision-language models by purifying initial representations before decoding.



Question: Please help me describe this image in detail.

LLAVA-1.5

The image features a skateboarder performing a trick on a ramp at a skate park. The skateboarder is in the middle of the trick, with the skateboard flipping upside down. Another person is sitting nearby, watching the skateboarder's impressive display. There are a few other objects in the scene, including a bottle located near the left side of the image, **and a backpack placed on the ground to the left of the skateboarder.** **Additionally, there is a chair situated in the background, slightly to the right of the skateboarder.**

w/ VTI

The image features a skateboarder performing a trick on a ramp at a skate park. The skateboarder is in the middle of the trick, with their skateboard flipping in the air. Another person is sitting nearby, watching the skateboarder's performance. **There are two bottles in the scene,** one located near the skateboarder **and the other further away.** **Additionally, there is a backpack placed on the ground, possibly belonging to one of the skateboarders.**

w/ VISTA

The image depicts a pool with two people in it. One person is on a skateboard, performing a trick in the pool while another person is sitting on a bench. The skateboarder is in the middle of a trick, and the other person is in the background. **There are two bottles in the scene,** one on the left side and **another on the right side.** The scene also features a skateboard and **a bench in the pool.**

w/ PTI (ours)

The image features a skateboarder performing a daring trick in mid-air, showcasing his skill and balance. The skateboarder is the main focus of the scene, with his skateboard clearly visible beneath him. In the background, there is another person who appears to be watching the skateboarder's performance. The scene takes place in a skate park, where skateboarders often practice and perform their tricks.

LLAVA-1.5

The image features a large gray cat sitting on a desk, occupying a significant portion of the space. **The cat is positioned near a computer keyboard,** which is placed on the desk. There are also several books scattered around the desk, with some placed near the cat and others further away. In addition to the cat and the books, **there is a TV in the background,** located on the right side of the desk. **A refrigerator can be seen in the far right corner of the image,** indicating that the desk is likely situated in a living or dining area.

w/ VTI

The image features a gray cat sitting on a desk, likely on a computer desk. **The cat is positioned near a computer keyboard** and appears to be looking at the camera. The desk also has a book and a cup placed on it. The cat seems to be relaxed and comfortable in its environment.",

w/ VISTA

The image features a cat sitting on a desk. The cat is in a relaxed position, and it appears to be a gray cat. **The desk is a workspace with a computer and a printer.** The cat is in a comfortable position on the desk, and it is a gray cat.

w/ PTI (ours)

The image features a large, fluffy gray cat sitting on a desk, occupying a significant portion of the scene. The cat appears to be looking directly at the camera, capturing the viewer's attention. The cat is positioned in the center of the image, with its body stretched out and filling the frame.



Question: Please help me describe this image in detail.

Figure 10. Qualitative examples of LLAVA-1.5. Hallucinated contents are marked in red.



Question: Please help me describe this image in detail.

Qwen-VL-Chat

In the image, a man is standing in a room holding a purple piece of luggage over his head. He appears to be struggling with the luggage, as it seems too large for him to lift easily. The man is standing in what seems to be an office or a similar professional setting. **The room has several chairs placed around the area,** along with a couple of fire extinguishers located in different spots for safety. **There is also a clock on the wall, which might serve as a timekeeping tool for the office.**

w/ VTI

The image features a man standing in a room with a piece of purple luggage on a tripod. The man appears to be holding the luggage or its handle, which is elevated since it is on a tripod. **The room has a few other pieces of furniture, including a chair towards the right side of the scene, and a bench in the far right corner. There is also a handbag placed on the floor near the left side of the room.**

w/ VISTA

A man is holding a purple case above his head. The man is wearing a gray shirt and blue jeans. The case appears to be a carrying case for a professional hair styling product. The man is in a room with a few other people, one of whom is a woman. A fire extinguisher is present in the room, **and there is a bench in the background.** A few other items are present in the scene, **including a couple of handbags and a book. A dining table is also visible in the room.** The man and other people are standing around the room, which appears to be a combination of a break room, a storage area, and a room for displaying a few items.

w/ PTI (ours)

In this image, a man is standing and holding a purple object above his head. The object appears to be a piece of luggage or a bag, and the man is holding it using a shoulder strap. **There are a few other people in the scene, but they are located further away and are not actively interacting with the man or the luggage. Overall, the focus of the image is on the man holding the bag and the strap above his head.**



Question: Please help me describe this image in detail.

Qwen-VL-Chat

The image features a woman sitting in a comfortable chair, **holding a cup in her hands.** She appears to be enjoying a relaxing moment, possibly sipping on a beverage while reading the newspaper. **The room has a vase placed on a side table near the chair, adding a touch of decoration to the space.** The woman is wearing jeans and a vest, and she is seated on a wooden rocking chair. The scene is a cozy one, with the woman taking some time for herself to relax and **enjoy her cup of tea or coffee** while reading the newspaper.

w/ VTI

The image features a woman sitting comfortably in a chair, holding a newspaper. She appears to be enjoying her time as she reads the newspaper while sitting in the chair. **The chair is positioned in a room, with a vase placed nearby on a small table. The table holds another vase,** adding a decorative touch to the scene. **The room also has a door, which is open,** and a window nearby. The window is covered by blinds, providing the woman with some shade or privacy as she reads her newspaper.

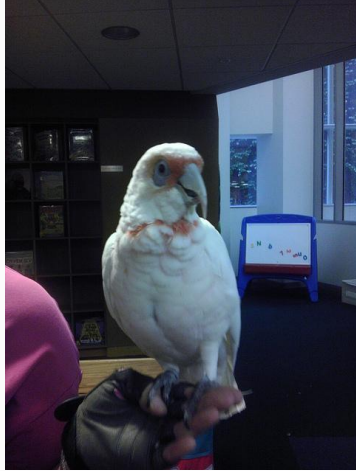
w/ VISTA

A woman is sitting in a chair with a newspaper. **She has a cup in her hand** and a flower in the room. **A clock is visible in the background.** A person is wearing a watch on their left hand. **A small dog is present in the scene, and a book is on the table.** A vase is placed on a side table near the chair. The room has a window with a blind and a curtain. The chair is on a carpet. The woman is wearing a vest and jeans. The image has a close-up view of a person in a chair.

w/ PTI (ours)

In the image, a woman is sitting in a comfortable chair, holding a newspaper. She appears relaxed and is smiling as she reads the paper. The chair she is sitting in is brown and appears to be a rocking chair, adding to the sense of relaxation. **Next to the woman, there is a small vase, possibly containing a flower, adding a touch of decoration to the scene.** The overall atmosphere seems tranquil, with the woman enjoying her time reading the newspaper in a cozy setting.

Figure 11. Qualitative examples of Qwen-VL-Chat. Hallucinated contents are marked in red.



Question: Please help me describe this image in detail.

DeepSeek-VL-Chat

In the image, there is a bird perched on a person's hand. The bird appears to be a cockatoo, characterized by its white feathers and distinctive red cheek patches. The person holding the bird is wearing a pink shirt. In the background, **there is a blue chair** and a whiteboard with some drawings on it. The setting seems to be an indoor environment, possibly an office or a classroom.

w/VTI

In this image, we see a white parrot perched on a person's hand. The parrot has a distinctive red patch on its head. The person is wearing a pink shirt. In the background, **there's a blue chair** and a whiteboard with some drawings on it. The setting appears to be an indoor room, possibly an office or a classroom, given the presence of the whiteboard.

w/VISTA

In the image, there is a white parrot with red patches around its neck and face, perched on someone's hand. The parrot has blue eyes and is looking straight ahead. The person holding the parrot is wearing a pink shirt and black gloves. In the background, **there's a blue plastic chair** and a whiteboard with some drawings on it. There are also some books on shelves in the background. The setting appears to be an office or a room with windows allowing natural light in.

w/PTI (ours)

In the image, a person is holding a white bird with a red patch on its neck. The bird is perched on the person's hand. The person is wearing a pink shirt. In the background, there is a blue object that appears to be a toy. The bird is the main focus of the image, and it is positioned in the center. The person's hand is visible at the bottom of the image, supporting the bird. The blue object in the background is located towards the right side of the image.



Question: Please help me describe this image in detail.

DeepSeek-VL-Chat

The image showcases a plate of food with two prominent items: a salad and a skewered meat. The salad is a mix of various greens, including arugula, and is accompanied by shredded carrots. The skewered meat appears to be grilled and has distinct grill marks. The plate is placed on a table, **and there's a fork visible on the right side of the plate**. The overall setting suggests a meal ready to be enjoyed.

w/VTI

The image predominantly showcases a plate with food items. Initially, I identified seven instances of carrots. Interestingly, upon a closer look, there's an apparent misclassification - some of the identified carrots actually correspond to celery, while several vegetables have been mistakenly classified as carrots. Upon a thorough analysis, I discovered a total of seventeen pieces of celery and **twenty-four pieces of broccoli**. All the vegetables are located within the bounding box.

w/VISTA

In the image, there are two meatballs placed on a plate. The meatballs are placed on a plate which also has a salad on it. The meatballs are brown in color and appear to be grilled. There are also some green leaves on the plate which could be part of the salad. There are also some carrots visible in the image. The plate is placed on a table which appears to be made of wood. **There are also some other objects like a chair** and a table visible in the image.

w/PTI (ours)

The image showcases a plate with two skewers of grilled meat, each skewer holding a piece of meat. The meat appears to be well-cooked with visible grill marks. Accompanying the meat on the plate is a fresh salad composed of arugula and shredded carrots. The salad provides a vibrant contrast to the grilled meat, with its green leaves and orange shreds. The plate is placed on a surface, and the overall presentation of the dish is appetizing.

Figure 12. Qualitative examples of DeepSeek-VL-Chat. Hallucinated contents are marked in red.