# Supplementary Material

## A. Observation and Analysis

In this section, we conduct additional experiments on LLaVA-1.5-13B [8], Janus-Pro-7B [3], and Qwen2.5-VL-7B [2] to verify that the similar characteristics observed for LLaVA-1.5-7B also hold for other LVLMs.

### A.1. Attention Pattern of Different LVLMs

We randomly selected 10,000 images from COCO2014 [7] and generated captions for these images using each of the three models with the prompt "Please describe the image in detail." Objects in the captions were annotated via ground truth annotations, yielding 22,362 true and 8,978 hallucinated objects for LLaVA-1.5-13B, 15,018 true and 2,924 hallucinated objects for Janus-Pro-7B, and 12,289 true and 2,687 hallucinated objects for Qwen2.5-VL-7B. We further computed each model's $\bar{\mathbf{A}}^r_{T_p}$, $\bar{\mathbf{A}}^h_{T_p}$, $\bar{\mathbf{A}}^r_V$, and $\bar{\mathbf{A}}^h_V$, visualized the results in Fig. 4-6.

The experimental results from all three models align with those from LLaVA-1.5-7B: during generation, true objects consistently assign higher average attention to both $V$ and $T_p$ than hallucinated objects. However, the disparity in attention over text tokens (typically on the order of 2x–3×) is substantially greater than that over image tokens (typically on the order of 1x–2×). This observation further supports that the knowledge compressed in $T_{\mathrm{p}}$ may facilitate more accurate predictions.

### A.2. Heads Behavior of Different LVLMs

For each of the three models (LLaVA-1.5-13B, Janus-Pro-7B, and Qwen2.5-VL-7B), we further computed their $\bar{\mathbf{A}}^r_{T_p}$, $\bar{\mathbf{A}}^h_{T_p}$, and $\mathcal{M}$, visualizing them in Fig. 1-3. Although $\bar{\mathbf{A}}^r_{T_p}$ consistently exceeds $\bar{\mathbf{A}}^h_{T_p}$, the magnitude of this ratio varies across different attention heads. For LLaVA-1.5-13B and Qwen2.5-VL-7B, some heads exhibit a 2× difference while others reach 5×; For Janus-Pro-7B, certain heads even display a 15× disparity. The inherent behavioral differences among attention heads highlight their distinct functional roles. Motivated by this observation, we propose AdaIAT to adaptively increase attention based on each head's specific characteristics.

## B. Extended Experiment results

### B.1. Comparison with More Methods

In this section, we compare attention intervention methods (PAI[9], HGAI[5], IAT, AdaIAT) with more methods (VCD[6], AGLA[1], LURE[12], OPERA[4]) on LLaVA-1.5-7B, evaluating their performance in mitigating hallucinations using the CHAIR metric, as well as their corresponding computational costs $t$. The experimental results are shown in the Table 1. It can be observed that VCD and AGLA, which involve contrastive decoding (each token generation requires two forward passes along with special processing of the input image), incur a time cost 3–4 times that of Greedy decoding. Similarly, OPERA, due to its rollback strategy, results in a time cost more than double that of Greedy decoding. In contrast, attention intervention methods achieve significantly better hallucination reduction while maintaining a computational time comparable to Greedy. Although LURE has the most aggressive hallucination suppression, its computational time is 9 times that of Greedy, making it impractical for real-time inference in real-world applications. This is because LURE employs a fine-tuned 13B LVLM as revisor to correct captions generated by the 7B model. On one hand, this introduces substantial computational costs for both fine-tuning and inference; on the other hand, such excessive revision leads to a sharp decline in F1. In comparison, attention intervention methods allow dynamic balancing between hallucination rate and accuracy by adjusting the amplification factor $\alpha$. Combined with their negligible additional computational overhead, attention intervention methods demonstrate strong practical applicability.

### B.2. Hyperparameter Ablation Analysis

Since the reduction of hallucination rates inevitably accompanies a decline in predictive performance or linguistic quality, blindly pursuing lower hallucination rates is undesirable. We randomly selected 500 images from the COCO dataset to generate captions and evaluated the hallucination rate via the $C_S$, $C_I$, prediction performance via F1 score, and language quality via $D_1$. We evaluated PAI, HGAI, IAT, and AdaIAT on LLaVA-1.5-7B, LLaVA-1.5-13B, Janus-Pro-7B, and Qwen2.5-VL-7B. Performance metrics under varying hyperparameter settings are reported in

Table 1. Comparison result of hallucination mitigation and time cost performance of different methods on LLaVA-1.5-7B

| Method | Greedy | VCD | AGLA | OPERA | LURE | PAI | HGAI | IAT | AdaIAT |
|---|---|---|---|---|---|---|---|---|---|
| $C_S \downarrow$ | 49.0 | 45.0 | 44.2 | 49.2 | 19.8 | 31.8 | 31.4 | 29.8 | 31.4 |
| $C_I \downarrow$ | 13.3 | 11.8 | 11.6 | 12.6 | 6.5 | 7.8 | 6.9 | 9.0 | 8.3 |
| F1 $\uparrow$ | 77.9 | 78.3 | 78.6 | 77.8 | 74.1 | 77.7 | 78.3 | 76.8 | 79.4 |
| $t(ms/token) \downarrow$ | 109.8 | 307.4 | 415.0 | 274.4 | 916.0 | 114.5 | 112.4 | 111.5 | 114.0 |

Tab. 3-10, thereby quantifying each method's trade-offs among hallucination rate, predictive capability, and language quality across different models. To ensure a fair comparison, we report in the main text the metrics corresponding to the lowest achievable hallucination rate while maintaining F1 without significant degradation.

### B.3. Experimental Results on OpenCHAIR

To further investigate the generalization capability of AdaIAT, this section presents full experimental results on OpenCHAIR. The CHAIR metric primarily relies on a closed vocabulary of objects (the 80 common object categories in MS-COCO), making it unable to detect open-vocabulary hallucinations (such as uncommon objects like "pearl" or "wheelchair"). Therefore, OpenCHAIR extends the CHAIR metric by relaxing its strong reliance on a closed vocabulary. It is achieved by utilizing LLaMA-2 [11] to rewrite MS-COCO descriptions, injecting diverse objects (e.g., "tricycle, corkscrew"), and then employing Stable Diffusion XL [10] to generate image-description pairs, which are subsequently curated by humans to ensure high quality. During evaluation, objects mentioned in captions are first parsed, and then an LLM determines whether each object appears in the ground-truth caption (labeling it as real, hallucinated, or uncertain), rather than relying on a fixed synonym dictionary. The hallucination rate is measured by $C_O$:

$$C_O = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{hallucinated objects and real objects}\}|}, \quad (1)$$

We selected 2,000 images from the OpenCHAIR dataset and generated captions using different LVLMs, prompted with "Please describe the image in detail." As shown in Tab. 2, AdaIAT achieves the lowest $C_O$ on both LLaVA-1.5-13B and Janus-Pro-7B, while maintaining text diversity comparable to the original Greedy decoding strategy. On Qwen2.5-VL, AdaIAT slightly underperforms PAI but still delivers a relatively strong performance. Given that PAI demonstrates degradation in language capabilities and weaker hallucination mitigation on other models (like LLaVA), AdaIAT achieves excellent hallucination mitigation across all evaluated models while preserving text diversity, achieving a superior trade-off.

Table 2. Extended OpenCHAIR results on different LVLMs, with the best values (excluding Greedy) highlighted in bold.

| Model | LLaVA-13B | | Janus-Pro | | Qwen2.5-VL | |
|---|---|---|---|---|---|---|
| Method | $C_O \downarrow$ | $D_1 \uparrow$ | $C_O \downarrow$ | $D_1 \uparrow$ | $C_O \downarrow$ | $D_1 \uparrow$ |
| Greedy | 0.283 | 0.62 | 0.327 | 0.63 | 0.349 | 0.64 |
| PAI | 0.276 | 0.58 | 0.321 | 0.63 | **0.303** | **0.67** |
| HGAI | 0.276 | 0.59 | 0.325 | 0.64 | 0.321 | 0.66 |
| IAT | 0.256 | **0.62** | 0.310 | **0.65** | 0.312 | 0.65 |
| AdaIAT | **0.235** | 0.61 | **0.298** | 0.64 | 0.315 | 0.66 |

## C. More Demo Cases

To more intuitively demonstrate the effectiveness of AdaIAT, we present additional demo cases in this section. Fig. 7-10 illustrate captions generated by three methods (PAI, HGAI, and AdaIAT) across four LVLMs. These cases highlight the repetitive descriptions induced by existing attention intervention methods (PAI and HGAI). As shown in Fig. 7-9, this phenomenon often occurs when the image contains few elements or only a single salient object. Due to insufficient attention to generated text tokens, PAI and HGAI forget the completed descriptions of objects and persistently describe the salient object, resulting in redundant repetitions. In contrast, AdaIAT retains memory of described content, terminates descriptions promptly, and maintains concise language without repetition. As shown in Fig. 10, with long-text-preferring models like Qwen2.5-VL, the model struggles to incorporate extensive preceding textual information due to insufficient attention, therefore losing logical language generation capabilities.

To demonstrate the hallucination mitigation effect of AdaIAT, Fig. 11-14 present demo cases of captions generated by Greedy, IAT, and AdaIAT methods across the above four models. Overall, IAT still occasionally exhibits hallucinations (Fig. 13 and Fig. 14), whereas AdaIAT delivers markedly superior hallucination mitigation. Even for models like Janus-Pro (Fig. 13) that exhibit extremely low hallucination levels, AdaIAT effectively eliminates sporadic, subtle hallucinations, showcasing its efficacy.

Table 3. Ablation analysis results for the amplification factor $\alpha$ on HGAI, PAI, and IAT in LLaVA-1.5-7B. Values highlighted in **bold** indicate those selected for reporting. The symbol "\" denotes that the model lost normal language generation capabilities, rendering metric evaluation infeasible.

| Method | HGAI | | | | PAI | | | | IAT | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\alpha$ | $C_S \downarrow$ | $C_I \downarrow$ | F1 $\uparrow$ | $D_1 \uparrow$ | $C_S \downarrow$ | $C_I \downarrow$ | F1 $\uparrow$ | $D_1 \uparrow$ | $C_S \downarrow$ | $C_I \downarrow$ | F1 $\uparrow$ | $D_1 \uparrow$ |
| 0.2 | 48.4 | 12.6 | 78.7 | 0.58 | 48.2 | 13.2 | 78.1 | 0.59 | 46.4 | 12.9 | 77.9 | 0.61 |
| 0.3 | 49.0 | 12.1 | 79.1 | 0.57 | 48.0 | 12.0 | 78.5 | 0.59 | 46.6 | 12.8 | 77.6 | 0.61 |
| 0.4 | 43.8 | 11.1 | 78.9 | 0.55 | 45.0 | 11.7 | 79.2 | 0.57 | 42.6 | 12.1 | 78.2 | 0.61 |
| 0.5 | **31.4** | **6.9** | **78.3** | **0.50** | **31.8** | **7.8** | **77.7** | **0.50** | 40.6 | 10.7 | 78.1 | 0.61 |
| 0.6 | 7.2 | 4.1 | 69.3 | 0.59 | 9.4 | 3.2 | 68.5 | 0.28 | 36.4 | 10.3 | 78.1 | 0.62 |
| 0.7 | 2.8 | 1.9 | 58.0 | 0.88 | 4.4 | 3.5 | 49.9 | 0.90 | 35.2 | 10.5 | 77.4 | 0.61 |
| 0.8 | 0.4 | 0.2 | 37.6 | 0.61 | 0.8 | 3.5 | 10.2 | 0.78 | **29.8** | **9.0** | **76.8** | **0.61** |
| 0.9 | \ | \ | \ | \ | \ | \ | \ | \ | 21.2 | 7.0 | 73.5 | 0.57 |

Table 4. Ablation analysis results of the balance coefficient $\beta$ and amplification factor $\alpha$ for AdaIAT on LLaVA-1.5-7B. Values highlighted in **bold** indicate those selected for reporting.

| $\beta$ | 0.25 | | | | 0.5 | | | | 1 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\alpha$ | $C_S \downarrow$ | $C_I \downarrow$ | F1 $\uparrow$ | $D_1 \uparrow$ | $C_S \downarrow$ | $C_I \downarrow$ | F1 $\uparrow$ | $D_1 \uparrow$ | $C_S \downarrow$ | $C_I \downarrow$ | F1 $\uparrow$ | $D_1 \uparrow$ |
| 3 | 36.0 | 9.5 | 79.3 | 0.61 | 36.2 | 9.5 | 79.0 | 0.61 | 35.8 | 10.0 | 78.9 | 0.60 |
| 4 | 35.2 | 9.2 | 79.2 | 0.61 | 34.6 | 9.5 | 78.9 | 0.60 | 36.2 | 10.1 | 78.5 | 0.60 |
| 5 | 32.8 | 8.9 | 79.2 | 0.60 | 33.8 | 9.3 | 78.4 | 0.60 | 35.2 | 9.6 | 78.0 | 0.60 |
| 6 | 32.4 | 8.9 | 78.7 | 0.60 | **31.4** | **8.3** | **79.4** | **0.60** | 32.8 | 8.9 | 78.6 | 0.60 |
| 7 | 31.8 | 9.4 | 77.7 | 0.60 | 31.8 | 8.6 | 78.7 | 0.59 | 32.6 | 8.8 | 77.8 | 0.60 |
| 8 | 31.4 | 9.1 | 77.9 | 0.59 | 29.4 | 7.9 | 79.0 | 0.59 | 30.4 | 8.4 | 78.5 | 0.59 |
| 9 | 29.4 | 8.3 | 78.5 | 0.59 | 30.0 | 7.8 | 78.3 | 0.58 | 27.4 | 8.0 | 77.9 | 0.59 |

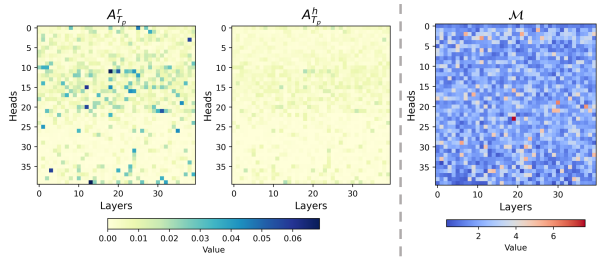



Figure 1. Left: Visualization of the average per-token attention weight to generated text for each layer and head of the LLaVA-1.5-13B, computed over the sets of real objects ($\bar{\mathbf{A}}^{r}_{T_p}$) and hallucinated objects ($\bar{\mathbf{A}}^{h}_{T_p}$). Right: The visualization of ratio $\bar{\mathbf{A}}^{r}_{T_p}/\bar{\mathbf{A}}^{h}_{T_p}$.

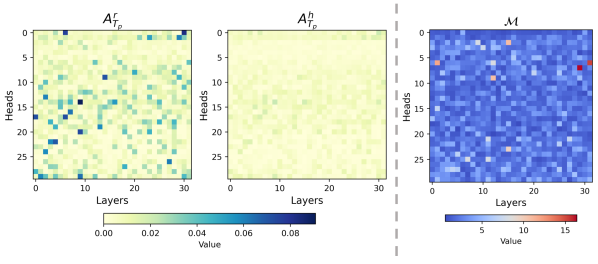



Figure 2. Left: Visualization of the average per-token attention weight to generated text for each layer and head of the Janus-Pro-7B, computed over the sets of real objects ($\bar{\mathbf{A}}^{r}_{T_p}$) and hallucinated objects ($\bar{\mathbf{A}}^{h}_{T_p}$). Right: The visualization of ratio $\bar{\mathbf{A}}^{r}_{T_p}/\bar{\mathbf{A}}^{h}_{T_p}$.

Table 5. Ablation analysis results for the amplification factor $\alpha$ on HGAI, PAI, and IAT in LLaVA-1.5-13B. Values highlighted in **bold** indicate those selected for reporting. The symbol "\" denotes that the model lost normal language generation capabilities, rendering metric evaluation infeasible.

| | HGAI | | | | PAI | | | | IAT | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\alpha$ | $C_S \downarrow$ | $C_I \downarrow$ | F1 $\uparrow$ | $D_1 \uparrow$ | $C_S \downarrow$ | $C_I \downarrow$ | F1 $\uparrow$ | $D_1 \uparrow$ | $C_S \downarrow$ | $C_I \downarrow$ | F1 $\uparrow$ | $D_1 \uparrow$ |
| 0.2 | 50.8 | 13.0 | 78.5 | 0.59 | 47.6 | 12.5 | 79.0 | 0.59 | 44.0 | 11.8 | 78.9 | 0.61 |
| 0.3 | 49.0 | 12.7 | 79.0 | 0.58 | 46.4 | 12.7 | 78.4 | 0.58 | 40.4 | 11.1 | 79.5 | 0.61 |
| 0.4 | 42.8 | 11.6 | 78.6 | 0.56 | 44.4 | 11.9 | 79.1 | 0.58 | 40.8 | 10.8 | 79.8 | 0.61 |
| 0.5 | **26.4** | **7.8** | **76.9** | **0.55** | **34.2** | **9.0** | **79.3** | **0.56** | 36.8 | 10.1 | 79.6 | 0.62 |
| 0.6 | 10.4 | 5.2 | 69.8 | 0.73 | 19.2 | 6.3 | 74.6 | 0.53 | 37.6 | 10.5 | 78.7 | 0.62 |
| 0.7 | 3.6 | 3.6 | 51.5 | 0.63 | 7.0 | 6.8 | 41.8 | 0.66 | 36.6 | 9.9 | 78.6 | 0.62 |
| 0.8 | 2.4 | 1.8 | 24.4 | 0.29 | \ | \ | \ | \ | **31.0** | **8.6** | **78.6** | **0.61** |
| 0.9 | \ | \ | \ | \ | \ | \ | \ | \ | 26.2 | 7.5 | 76.7 | 0.55 |

Table 6. Ablation analysis results of the balance coefficient $\beta$ and amplification factor $\alpha$ for AdaIAT on LLaVA-1.5-13B. Values highlighted in **bold** indicate those selected for reporting.

| $\beta$ | 0.25 | | | | 0.5 | | | | 1 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\alpha$ | $C_S \downarrow$ | $C_I \downarrow$ | F1 $\uparrow$ | $D_1 \uparrow$ | $C_S \downarrow$ | $C_I \downarrow$ | F1 $\uparrow$ | $D_1 \uparrow$ | $C_S \downarrow$ | $C_I \downarrow$ | F1 $\uparrow$ | $D_1 \uparrow$ |
| 9 | 32.4 | 8.7 | 78.3 | 0.61 | 30.2 | 8.0 | 78.7 | 0.61 | 30.4 | 7.9 | 78.9 | 0.61 |
| 10 | 31.6 | 8.2 | 78.9 | 0.60 | 28.4 | 7.3 | 79.5 | 0.60 | 28.8 | 7.4 | 78.8 | 0.60 |
| 11 | 31.0 | 8.2 | 78.0 | 0.60 | 28.8 | 8.1 | 78.8 | 0.60 | 26.6 | 7.4 | 79.0 | 0.60 |
| 12 | 30.2 | 7.6 | 78.0 | 0.60 | 26.8 | 6.9 | 79.3 | 0.60 | **25.2** | **6.1** | **79.2** | **0.60** |
| 13 | 29.0 | 7.6 | 78.0 | 0.60 | 28.0 | 7.3 | 77.7 | 0.59 | 25.8 | 7.3 | 78.4 | 0.60 |
| 14 | 27.8 | 7.8 | 78.3 | 0.60 | 27.4 | 8.0 | 77.9 | 0.59 | 24.2 | 6.9 | 78.7 | 0.60 |
| 15 | 26.6 | 6.9 | 78.1 | 0.59 | 23.2 | 6.7 | 78.1 | 0.59 | 23.8 | 8.7 | 78.3 | 0.59 |

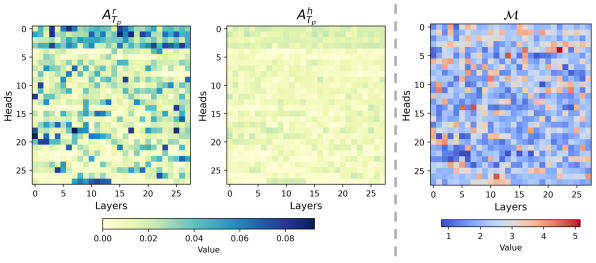
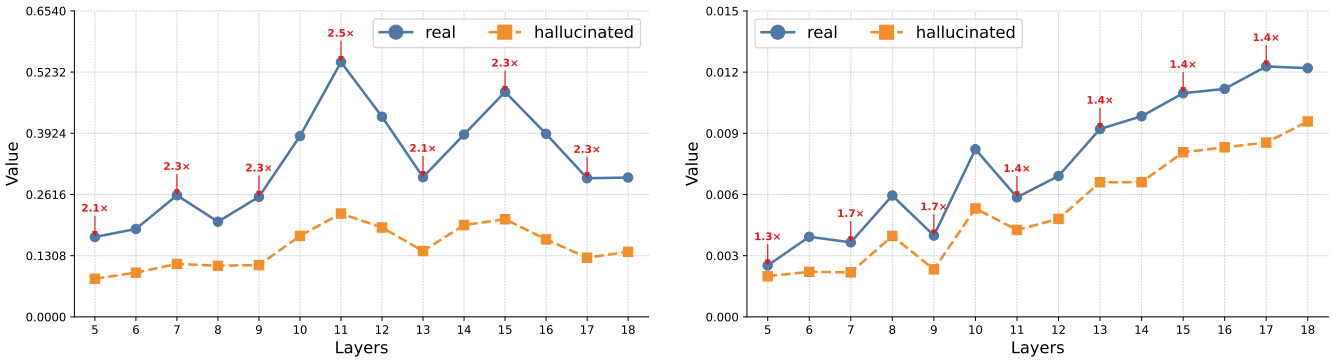


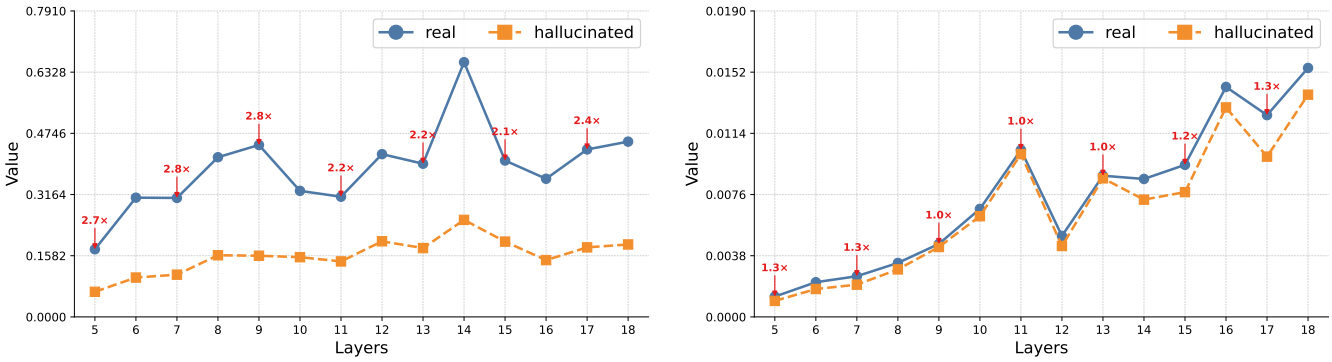

Figure 3. Left: Visualization of the average per-token attention weight to generated text for each layer and head of the Qwen2.5-VL-7B, computed over the sets of real objects ($\bar{\mathbf{A}}^{r}_{T_p}$) and hallucinated objects ($\bar{\mathbf{A}}^{h}_{T_p}$). Right: The visualization of ratio $\bar{\mathbf{A}}^{r}_{T_p}/\bar{\mathbf{A}}^{h}_{T_p}$.




(a) $\bar{\mathbf{A}}^{r}_{T_p}$ and $\bar{\mathbf{A}}^{h}_{T_p}$ (Layers 5-18)

(b) $\bar{\mathbf{A}}^{r}_{V}$ and $\bar{\mathbf{A}}^{h}_{V}$ (Layers 5-18)

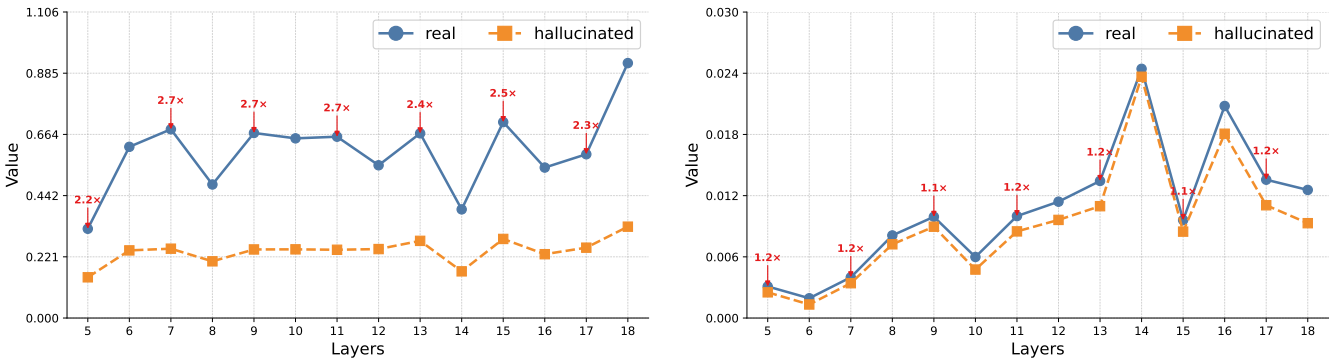
Figure 4. Visualization of the average per-token attention weights from $t_{n+1}$ to $T_p$ ($\bar{\mathbf{A}}^{r}_{T_p}$ and $\bar{\mathbf{A}}^{h}_{T_p}$) and to $V$ ($\bar{\mathbf{A}}^{r}_{V}$ and $\bar{\mathbf{A}}^{h}_{V}$) for LLaVA-1.5-13B, showing only layers 5–18 for clearer observation.




(a) $\bar{\mathbf{A}}^{r}_{T_p}$ and $\bar{\mathbf{A}}^{h}_{T_p}$ (Layers 5-18)

(b) $\bar{\mathbf{A}}^{r}_{V}$ and $\bar{\mathbf{A}}^{h}_{V}$ (Layers 5-18)

Figure 5. Visualization of the average per-token attention weights from $t_{n+1}$ to $T_p$ ($\bar{\mathbf{A}}^{r}_{T_p}$ and $\bar{\mathbf{A}}^{h}_{T_p}$) and to $V$ ($\bar{\mathbf{A}}^{r}_{V}$ and $\bar{\mathbf{A}}^{h}_{V}$) for Janus-Pro-7B, showing only layers 5–18 for clearer observation.




(a) $\bar{\mathbf{A}}^{r}_{T_p}$ and $\bar{\mathbf{A}}^{h}_{T_p}$ (Layers 5-18)

(b) $\bar{\mathbf{A}}^{r}_{V}$ and $\bar{\mathbf{A}}^{h}_{V}$ (Layers 5-18)

Figure 6. Visualization of the average per-token attention weights from $t_{n+1}$ to $T_p$ ($\bar{\mathbf{A}}^{r}_{T_p}$ and $\bar{\mathbf{A}}^{h}_{T_p}$) and to $V$ ($\bar{\mathbf{A}}^{r}_{V}$ and $\bar{\mathbf{A}}^{h}_{V}$) for Qwen2.5-VL-7B, showing only layers 5–18 for clearer observation.

Table 7. Ablation analysis results for the amplification factor $\alpha$ on HGAI, PAI, and IAT in Janus-Pro-7B. Values highlighted in **bold** indicate those selected for reporting. The symbol "\" denotes that the model lost normal language generation capabilities, rendering metric evaluation infeasible.

| | HGAI | | | | PAI | | | | IAT | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\alpha$ | $C_S \downarrow$ | $C_I \downarrow$ | F1 $\uparrow$ | $D_1 \uparrow$ | $C_S \downarrow$ | $C_I \downarrow$ | F1 $\uparrow$ | $D_1 \uparrow$ | $C_S \downarrow$ | $C_I \downarrow$ | F1 $\uparrow$ | $D_1 \uparrow$ |
| 0.1 | 26.6 | 6.9 | 77.0 | 0.62 | 26.6 | 6.8 | 77.3 | 0.62 | 25.8 | 6.7 | 76.5 | 0.62 |
| 0.2 | 28.0 | 7.5 | 77.0 | 0.62 | 27.4 | 7.0 | 76.9 | 0.62 | 26.2 | 6.6 | 76.6 | 0.63 |
| 0.3 | 26.4 | 6.7 | 76.3 | 0.62 | 27.2 | 6.9 | 77.3 | 0.61 | 22.6 | 6.0 | 76.7 | 0.63 |
| 0.4 | **21.0** | **5.3** | **75.9** | **0.62** | **20.4** | **5.6** | **76.1** | **0.61** | 22.8 | 6.0 | 76.3 | 0.63 |
| 0.5 | 18.0 | 6.0 | 68.0 | 0.60 | 11.6 | 4.4 | 67.7 | 0.61 | 21.2 | 5.7 | 76.1 | 0.64 |
| 0.6 | 11.8 | 3.4 | 57.4 | 0.43 | 1.2 | 1.4 | 26.5 | 0.62 | **20.6** | **5.3** | **75.3** | **0.65** |
| 0.7 | 8.8 | 3.1 | 51.6 | 0.23 | \ | \ | \ | \ | 17.4 | 5.1 | 73.8 | 0.65 |
| 0.8 | 6.6 | 6.1 | 34.5 | 0.18 | \ | \ | \ | \ | 13.2 | 8.9 | 69.0 | 0.64 |

Table 8. Ablation analysis results of the balance coefficient $\beta$ and amplification factor $\alpha$ for AdaIAT on Janus-Pro-7B. Values highlighted in **bold** indicate those selected for reporting.

| $\beta$ | 0.25 | | | | 0.5 | | | | 1 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\alpha$ | $C_S \downarrow$ | $C_I \downarrow$ | F1 $\uparrow$ | $D_1 \uparrow$ | $C_S \downarrow$ | $C_I \downarrow$ | F1 $\uparrow$ | $D_1 \uparrow$ | $C_S \downarrow$ | $C_I \downarrow$ | F1 $\uparrow$ | $D_1 \uparrow$ |
| 1 | 24.8 | 6.4 | 76.3 | 0.63 | 24.8 | 6.2 | 75.8 | 0.64 | 25.2 | 6.5 | 76.2 | 0.63 |
| 2 | 21.8 | 5.9 | 75.7 | 0.64 | 22.8 | 6.1 | 75.3 | 0.64 | 22.0 | 6.1 | 76.3 | 0.64 |
| 3 | 20.0 | 6.0 | 76.2 | 0.64 | 21.4 | 6.0 | 75.4 | 0.64 | 21.6 | 5.8 | 75.9 | 0.64 |
| 4 | 17.6 | 4.9 | 75.7 | 0.64 | 20.8 | 5.7 | 75.1 | 0.64 | **19.0** | **4.9** | **76.5** | **0.64** |
| 5 | 19.4 | 5.1 | 75.2 | 0.65 | 20.4 | 5.5 | 75.0 | 0.64 | 17.8 | 4.6 | 75.4 | 0.65 |
| 6 | 16.2 | 4.4 | 74.9 | 0.64 | 18.0 | 4.7 | 75.1 | 0.65 | 18.0 | 5.7 | 74.2 | 0.65 |
| 7 | 16.6 | 4.8 | 73.8 | 0.65 | 19.0 | 5.1 | 74.3 | 0.65 | 18.2 | 5.6 | 74.1 | 0.64 |

Table 9. Ablation analysis results for the amplification factor $\alpha$ on HGAI, PAI, and IAT in Qwen2.5-VL-7B. Values highlighted in **bold** indicate those selected for reporting.

| | HGAI | | | | PAI | | | | IAT | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\alpha$ | $C_S \downarrow$ | $C_I \downarrow$ | F1 $\uparrow$ | $D_1 \uparrow$ | $C_S \downarrow$ | $C_I \downarrow$ | F1 $\uparrow$ | $D_1 \uparrow$ | $C_S \downarrow$ | $C_I \downarrow$ | F1 $\uparrow$ | $D_1 \uparrow$ |
| 0.1 | 36.6 | 8.5 | 77.0 | 0.65 | 37.0 | 9.0 | 75.7 | 0.65 | 32.0 | 8.0 | 75.9 | 0.66 |
| 0.2 | 34.4 | 8.1 | 76.9 | 0.65 | 38.4 | 10.2 | 74.9 | 0.65 | **29.2** | **7.7** | **76.4** | **0.67** |
| 0.3 | 33.8 | 8.2 | 77.3 | 0.64 | 37.0 | 9.2 | 74.6 | 0.64 | 23.6 | 6.9 | 76.0 | 0.67 |
| 0.4 | **30.4** | **8.0** | **76.8** | **0.65** | 37.2 | 9.0 | 73.6 | 0.63 | 25.2 | 6.9 | 75.8 | 0.69 |
| 0.5 | 30.4 | 7.4 | 76.0 | 0.65 | **32.0** | **8.2** | **74.7** | **0.64** | 24.0 | 7.0 | 74.1 | 0.70 |
| 0.6 | 27.0 | 8.2 | 73.9 | 0.66 | 24.8 | 6.9 | 69.9 | 0.65 | 20.8 | 6.0 | 72.6 | 0.70 |
| 0.7 | 16.4 | 6.2 | 67.9 | 0.76 | 9.4 | 4.5 | 58.6 | 0.79 | 18.0 | 6.0 | 71.4 | 0.71 |
| 0.8 | 6.6 | 4.1 | 57.4 | 0.85 | 3.6 | 2.2 | 52.5 | 0.87 | 18.6 | 5.8 | 70.5 | 0.70 |

Table 10. Ablation analysis results of the balance coefficient $\beta$ and amplification factor $\alpha$ for AdaIAT on Qwen2.5-VL. Values highlighted in **bold** indicate those selected for reporting.

| $\beta$ | 0.25 | | | | 0.5 | | | | 1 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\alpha$ | $C_S \downarrow$ | $C_I \downarrow$ | F1 $\uparrow$ | $D_1 \uparrow$ | $C_S \downarrow$ | $C_I \downarrow$ | F1 $\uparrow$ | $D_1 \uparrow$ | $C_S \downarrow$ | $C_I \downarrow$ | F1 $\uparrow$ | $D_1 \uparrow$ |
| 6 | 31.2 | 7.3 | 76.7 | 0.66 | 30.6 | 7.3 | 77.0 | 0.66 | 31.2 | 7.6 | 76.8 | 0.66 |
| 7 | 31.2 | 7.4 | 76.2 | 0.66 | 32.8 | 7.8 | 76.6 | 0.66 | 30.8 | 7.7 | 76.8 | 0.66 |
| 8 | 31.6 | 7.6 | 76.9 | 0.66 | 30.2 | 7.5 | 76.8 | 0.66 | 31.4 | 7.4 | 77.1 | 0.66 |
| 9 | **28.4** | **6.9** | **76.8** | **0.66** | 31.2 | 7.7 | 76.7 | 0.66 | 30.2 | 7.4 | 76.7 | 0.66 |
| 10 | 30.6 | 7.6 | 76.2 | 0.66 | 33.6 | 8.4 | 75.9 | 0.66 | 29.4 | 7.6 | 76.9 | 0.66 |
| 11 | 29.8 | 7.2 | 76.9 | 0.66 | 29.6 | 7.1 | 76.9 | 0.66 | 30.6 | 7.6 | 77.1 | 0.66 |
| 12 | 30.4 | 7.4 | 76.3 | 0.66 | 30.6 | 7.8 | 76.3 | 0.66 | 30.2 | 7.4 | 77.2 | 0.66 |

Please describe the image in detail.

**PAI:** The image features a group of zebras standing together in a dirt field. There are five zebras in the scene, with one of them being the most prominent. The zebras are standing in a group, with some of them being closer to the camera and others further away. The zebras are standing in a dirt field, with some of them being closer to the camera and others further away. The zebras are standing in a group, with some of them being closer to the camera and others further away……

**HGAI:** The image features a group of zebras in a zoo enclosure, with some of them standing close to a fence. There are five zebras in total, with one of them being the most prominent. The zebras are standing in a dirt area, and they are all looking at the camera. The zebras are standing in a dirt field, and they are all looking at the camera. The zebras are standing in a dirt field, and they are all looking at the camera.

**AdaIAT**: The image features a group of zebras standing together in a dirt field. There are five zebras in total, with some standing closer to the foreground and others further back in the scene. The zebras are standing behind a fence, which separates them from the viewer. The fence is visible in the background, and the zebras appear to be enjoying their time together in the field.

Please describe the image in detail.

**PAI:** The image features a large brick building with a prominent clock tower on top of it. The clock tower is situated on the side of the building, making it a distinctive and unique architectural feature. The building is located in a city, surrounded by other buildings, which are also made of brick. The clock tower stands out among the other buildings, making it a focal point in the scene.

**HGAI:** The image features a large clock tower with two clocks on its side, located on top of a building. The clock tower is situated in the middle of a city, surrounded by tall buildings. The clock tower is situated in the middle of a city, and it is surrounded by tall buildings. The clock tower is a prominent landmark in the city, and it is a notable feature of the urban landscape.

**AdaIAT**: The image features a large brick building with a prominent clock tower on top. The clock is situated in the middle of the tower, making it easily visible from a distance. The building is located in a city, surrounded by other tall buildings. There are two clocks on the tower, one larger and more prominent, and another smaller one. The building appears to be a prominent landmark in the city, with its clock tower standing out among the surrounding structures.

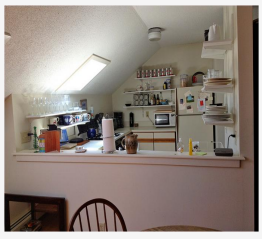
Figure 7. Demo cases of repetitive descriptions phenomenon in LLaVA-1.5-7B, with repeated segments highlighted in **blue**.

Please describe the image in detail.

**PAI:** The image features a cow with a large hole in its side, which appears to be a unique and unusual feature. The cow is standing in a barn, and its backside is facing the camera. The cow's skin appears to be scarred, possibly due to the hole in its side. The cow's skin appears to be scarred, which might indicate that it has experienced some form of injury or trauma. The cow's skin appears to be scarred, possibly due to the hole in its side.

**HGAI:** The image features a group of zebras in a zoo enclosure, with some of them standing close to a fence. There are five zebras in total, with one of them being the most prominent. The zebras are standing in a dirt area, and they are all looking at the camera. The zebras are standing in a dirt field, and they are all looking at the camera. The zebras are standing in a dirt field, and they are all looking at the camera.

**AdaIAT**: The image features a group of zebras standing together in a dirt field. There are five zebras in total, with some standing closer to the foreground and others further back in the scene. The zebras are standing behind a fence, which separates them from the viewer. The fence is visible in the background, and the zebras appear to be enjoying their time together in the field.



Please describe the image in detail.

**PAI:** The image shows a cozy kitchen or kitchen area with a slanted ceiling. The space is well-lit with a skylight on the slanted ceiling. There are various kitchen items and appliances visible, including a microwave, a toaster, and a refrigerator. There are also shelves and a counter space with a chair in front of it.

**HGAI:** The image depicts a kitchen area with a white countertop and a dining table. The kitchen is equipped with various appliances, including a microwave, oven, and refrigerator. There are also several cups, bowls, and bottles scattered throughout the space. The kitchen is well-organized and clean, with a variety of items such as cups, bowls, and bottles. In addition to the kitchen appliances, there are also several cups, bowls, and bottles placed in different areas of the kitchen.

**AdaIAT**: The image features a small kitchen area with a white refrigerator and a microwave oven. The kitchen is equipped with various appliances, including a toaster oven, a blender, and a coffee maker. There are also several cups and bottles placed around the kitchen, as well as a bowl and a spoon.\n\nIn addition to the kitchen appliances, there is a dining table with a chair positioned near it. The table is surrounded by chairs, and there is a book placed on the table.

Figure 8. Demo cases of repetitive descriptions phenomenon in LLaVA-1.5-13B, with repeated segments highlighted in **blue**.

Please describe the image in detail.

**PAI:** The image shows a hot dog in a white, rectangular, disposable foam takeout container. The hot dog is in a bun and is topped with a pickle spear, a pickle spear, a pickle spear, a pickle spear, a pickle spear, a pickle spear, a pickle spear, a pickle spear……

**HGAI:** The image shows a hot dog in a white, disposable, and open-topped container. The hot dog is placed in a white, rectangular, and slightly open container. The hot dog is topped with a variety of condiments and garnishes, including a greenish-yellowish sausage, a red-orange hair, and a blue-ish background.

**AdaIAT**: The image shows a hot dog in a white, foam takeout container. The hot dog is in a bun and topped with various ingredients. There is a pickle spear, sliced cucumber, tomato slices, and what appears to be mustard or another sauce. The hot dog is placed on a dark wooden surface, and there is a piece of paper or a napkin partially visible on the right side of the image.

Please describe the image in detail.

**PAI:** The image features a large brick building with a prominent clock tower on top of it. The clock tower is situated on the side of the building, making it a distinctive and unique architectural feature. The building is located in a city, surrounded by other buildings, which are also made of brick. The clock tower stands out among the other buildings, making it a focal point in the scene.

**HGAI:** The image shows a bottle of red wine and a glass of red wine placed on a glass table. The bottle of red wine has a label that reads \"Cupcake\" and \"Pinot Noir\". The bottle is placed next to a glass of red wine on a wine glass. The bottle is labeled as \"Pinot Noir\". The bottle is placed on a reflective glass table. The wine bottle and the glass of red wine are both placed on the table. The bottle is green and the wine glass is filled with red wine. The bottle has a yellow label with a decorative design. The wine glass is filled with red wine. The bottle is placed next to a glass of red wine on the table. The wine bottle is green and has a yellow label with a decorative design. The wine glass is filled with red wine. The bottle is labeled as \"Cupcake\". The wine bottle is placed on the table next to the glass of red wine. The wine bottle is green and has a yellow label with a decorative design. The wine glass is filled with red wine. The wine bottle is placed next to the glass of red wine on the table. The wine bottle is labeled as \"Cupcake\".

**AdaIAT**: The image features a large brick building with a prominent clock tower on top. The clock is situated in the middle of the tower, making it easily visible from a distance. The building is located in a city, surrounded by other tall buildings. There are two clocks on the tower, one larger and more prominent, and another smaller one. The building appears to be a prominent landmark in the city, with its clock tower standing out among the surrounding structures.

Figure 9. Demo cases of repetitive descriptions phenomenon in Janus-Pro-7B, with repeated segments highlighted in **blue**.

Please describe the image in detail.

**PAI:** The image showcases a buffet-style meal spread across a table, offering a variety of dishes for guests to choose from. The table is laden with an assortment of foods, including: 1. Salads and Fresh Vegetables: A large platter of mixed greens topped with cherry tomatoes, cucumbers, and other vegetables. A bowl of mixed greens dressed with olive oil and vinegar. 2. Starchy Side Dishes: A bowl of corn kernels. A bowl of mixed vegetables (carrots, peas, corn). A bowl of green beans. 3. Main Courses: A casserole dish filled with a creamy mixture, possibly a pasta bake or a casserole with a crusty topping. A tray of mixed vegetables (carrots, peas, corn). 4. Desserts: A bowl of mixed berries (strawberries, raspberries) with a dollop of whipped cream on top. A bowl of mixed berries (strawberries, raspberries) with a dollop of whipped cream on top. A bowl of mixed berries (strawberries, raspberries) with a dollop of whipped cream on top.....

**AdaIAT**: The image shows a table filled with an assortment of food, suggesting a buffet-style meal or gathering. The table is covered with various dishes, including:
1. Salads and Vegetables: There are several bowls containing fresh salads, mixed greens, and other vegetables like carrots and peas. 2. Main Dishes: A large casserole dish appears to contain a baked dish, possibly a casserole or gratin, with a golden-brown crust on top. 3. Side Dishes: Bowls of corn kernels, mashed potatoes, and what seems to be a creamy pasta dish are visible. 4. Pies and Pastries: On the left side, there are pies or tarts in metal pans, along with some bread rolls or pastries. 5. Other Dishes: There are also plates with what looks like sliced tomatoes and other toppings, possibly for sandwiches or open-faced sandwiches.
The setting suggests a casual dining environment, such as a family gathering, a potluck, or a community event. The variety of dishes indicates that the meal is meant to cater to different tastes and preferences.

Please describe the image in detail.

**HGAI:** The image shows a kitchen workspace with various ingredients and tools laid out on a wooden table, likely for baking or cooking. Here's a detailed breakdown of the items:
1. Ingredients: Bananas: Two ripe bananas. Eggs: Three eggs. Flour: A bag of white flour (likely all-purpose). Sugar: A bag of sugar. Salt: A small container of salt. Baking Powder: A box of baking powder. Baking Soda: A box of baking soda. Milk: A carton of milk. Oil: A bottle of oil. 2. Tools/Equipment: A laptop is open and placed in the background. A metal tray holds the bananas and an orange. A measuring cup and a baking pan are visible. A bamboo mat is used as a base for some items. 3. Other Items: A pack of garlic oil. A pack of what appears to be a spice or seasoning mix. A pack of what might be a flavoring or seasoning mix. A pack of what looks like a sweetener or sugar substitute. A pack of what could be a flavoring or seasoning mix. A pack of what seems to be a sweetener or sugar substitute. A pack of what might be a flavoring or seasoning mix……

**AdaIAT**: The image shows a kitchen countertop with various ingredients and items laid out, likely for baking or cooking. Here's a detailed description:
1. Bananas: There are three ripe bananas on a metal tray. 2. Lemon: A whole lemon is placed next to the bananas. 3. Measuring Cup: A clear plastic measuring cup is visible on the left side of the image. 4. Eggs: Three eggs are in a small aluminum foil container. 5. Flour: A large bag of white flour labeled in Japanese characters is prominently displayed. 6. Baking Ingredients: Other ingredients include a carton of milk, a box of baking powder (with Japanese text), a packet of salt, and a bottle of oil or vinegar. 7. Laptop: A silver laptop is open on the table, suggesting that someone might be following a recipe online.
The setup indicates preparation for a baking project, possibly a cake or bread, given the presence of bananas, eggs, and flour. The laptop suggests that the recipe might have been sourced from an online recipe or tutorial.

Figure 10. Demo cases of repetitive descriptions phenomenon in Qwen2.5-VL-7B, with repeated segments highlighted in **blue**.

Please describe the image in detail.

**Greedy:** The image features a man standing on a wooden platform, observing two giraffes in a zoo enclosure. The man is wearing an orange shirt and appears to be enjoying his time at the zoo. The giraffes are positioned close to each other, with one being slightly larger than the other.\n\nThere are several other people in the scene, some of them standing near the man and others scattered around the platform. A bench can be seen in the background, providing a place for visitors to sit and enjoy the view of the giraffes.

**IAT**: The image features a man standing on a wooden deck, observing two giraffes in a zoo enclosure. The giraffes are positioned close to each other, with one giraffe being closer to the man and the other giraffe being further away. The man appears to be enjoying his time at the zoo, watching the giraffes as they move around their enclosure.

**AdaIAT**: The image features a man standing on a wooden platform, observing two giraffes in their enclosure. The man is wearing a hat and appears to be enjoying the view of the giraffes. The giraffes are standing close to each other, with one being closer to the man and the other further away.\n\nIn the background, there are several other people present, likely observing the giraffes as well. The scene suggests a zoo or wildlife park where visitors can observe the animals up close.

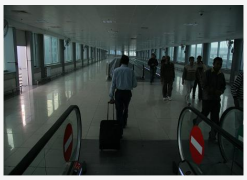
Please describe the image in detail.

**Greedy:** The image captures a snowboarder in mid-air, performing a jump off a ramp. The snowboarder is in the center of the scene, showcasing their impressive skills. The ramp is located in the foreground, with the snowboarder soaring above it. In the background, there are two people standing, possibly watching the snowboarder's performance or waiting for their turn to ride. The scene also features a truck parked in the background, possibly used for transporting snowboarding equipment or providing support for the event.

**AdaIAT**: The image captures a snowboarder in mid-air, performing a jump off a ramp. The snowboarder is in the middle of the jump, showcasing their skill and athleticism. The scene is set in a snowy environment, with the ramp visible in the background. The snowboarder is the main focus of the image, with their snowboard clearly visible as they soar through the air.

**AdaIAT**: The image captures a snowboarder in mid-air, performing a jump off a ramp. The snowboarder is in the middle of the jump, showcasing their skill and athleticism. The scene is set in a snowy environment, with the ramp visible in the background. The snowboarder is the main focus of the image, with their snowboard clearly visible as they soar through the air.

Figure 11. Qualitative results of hallucination mitigation in LLaVA-1.5-7B,, with hallucinated segments highlighted in **red**.

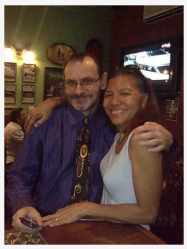


Please describe the image in detail.

**Greedy:** The image depicts a busy airport terminal with several people walking around. A man is pulling a suitcase behind him as he walks down the hallway, likely heading towards his destination. There are other people in the scene, some of them carrying handbags and backpacks. The airport terminal has a large open area with a few chairs placed around the space. A TV can be seen mounted on the wall, providing information or entertainment for the travelers. The overall atmosphere of the terminal is bustling, with people moving about and preparing for their flights.

**IAT**: The image depicts a busy airport terminal with several people walking around. One man is pulling a suitcase, likely preparing to board a plane or disembark from one. The people in the scene are engaged in various activities, such as walking, talking, or waiting for their flights.\n\nThe airport terminal is filled with people, some of whom are carrying luggage, while others are simply walking around. The scene captures the hustle and bustle of an airport terminal, with people going about their daily routines.

**AdaIAT**: The image depicts a busy airport terminal with several people walking around. A man is pulling a suitcase behind him as he walks through the terminal. There are other people in the scene, some of whom are also carrying luggage.\n\nIn addition to the man with the suitcase, there are at least six other people visible in the image. Some aisles, and some of them are carrying backpacks. The scene captures the hustle and bustle of an airport terminal, with people moving about and attending to their travel needs.



Please describe the image in detail.

**Greedy:** The image features a man and a woman standing close to each other, hugging each other in a friendly manner. They are both smiling, enjoying each other's company. The man is wearing a tie, which adds a touch of formality to the scene.\n\nIn the background, there is a TV mounted on the wall, and a cell phone can be seen placed on a surface nearby. Additionally, there are two chairs in the room, one on the left side and another on the right side. Another person is visible in the background, but they are not the main focus of the scene.

**IAT**: The image features a man and a woman standing close to each other, possibly hugging or posing for a picture. They are both wearing ties, suggesting a formal or semi-formal occasion. \n\nIn the background, there is a TV mounted on the wall, and a cell phone can be seen placed on a surface. The scene appears to be set in a bar or a similar social gathering venue.

**AdaIAT**: The image features a man and a woman standing close to each other, hugging each other in a warm and friendly manner. They are both smiling, indicating a positive and happy atmosphere between them.\n\nIn the background, there is a TV mounted on the wall, which adds to the cozy ambiance of the scene. The man and woman appear to be enjoying each other's company, creating a pleasant and intimate moment.

Figure 12. Qualitative results of hallucination mitigation in LLaVA-1.5-13B, with hallucinated segments highlighted in **red**.

Please describe the image in detail.

**Greedy:** The image shows a person wearing a maroon hoodie with a graphic design on the front. The person is standing in a kitchen, holding a pizza cutter in their right hand and cutting a pizza on a metal tray. The pizza appears to have toppings such as green peppers and possibly onions. The kitchen has a green countertop, and there are some containers and a bottle visible in the background. The person is wearing brown pants and has their left hand resting on the counter.

**IAT**: The image shows a person wearing a maroon hoodie with a graphic design on the front. The person is cutting a pizza with a pizza cutter on a green countertop. The pizza appears to have toppings such as green peppers. In the background, there are some kitchen appliances and items visible, including a microwave and some containers.

**AdaIAT**: The image shows a person wearing a maroon hoodie with a graphic design on the front. The person is standing in a kitchen, holding a pizza cutter and cutting into a pizza on a metal tray. The kitchen has a green countertop and some kitchen appliances in the background. The person is focused on cutting the pizza.

Please describe the image in detail.

**Greedy:** The image shows a person standing in a living room, holding a white object that appears to be a Wii remote. The person is wearing a black sweater, a green shirt, and brown pants. There is another person in the background, standing near a table with some items on it. The room has a red couch, a framed picture on the wall, and a doorway leading to another room.",

**IAT**: The image shows two people in a living room setting. The person in the foreground is holding a white object, which appears to be a Wii Remote, commonly used for playing video games on the Nintendo Wii console. The person is wearing a black sweater and has a flower accessory in their hair. The background features a red couch with a remote control and a blue cushion on it. There is a framed picture on the wall, and a doorway leading to another room. The overall setting appears to be a cozy and casual living room.

**AdaIAT**: The image shows a person standing in a living room, holding a white object that appears to be a Wii Remote, commonly used for playing video games on the Nintendo Wii console. The person is wearing a black sweater over a green shirt and has a flower accessory in their hair. In the background, there is another person standing, and there is a red couch with a remote control on it. The room has a modern and minimalistic decor, with a framed picture on the wall and a small table with some items on it.

Figure 13. Qualitative results of hallucination mitigation in Janus-Pro-7B, with hallucinated segments highlighted in **red**.

Please describe the image in detail.

**Greedy:** The image shows a bunch of bananas placed on a round, light-colored plate. The bananas appear ripe with some brown spots, indicating they are ready to eat or have been slightly overripe. One banana is lying horizontally across the others, and it has more pronounced brown spots compared to the others. The bananas are still attached at the stem, which is green and unripe-looking.\n\nIn the background, there is a kitchen countertop with various items. To the left, there is a small white measuring cup, a glass jar, and what appears to be a bottle of beer. On the right side, there is a knife block containing several knives, and part of a black object, possibly a cutting board or another kitchen item, is visible. The lighting in the image is warm, giving it a cozy, homey feel.

**IAT**: The image shows a bunch of bananas placed on a wooden cutting board. The bananas have some brown spots, indicating they are ripe. There is a sticker with a blue logo on one of the bananas. In the background, there are various kitchen items, including a glass container and what appears to be a spice rack or utensil holder. The setting seems to be a kitchen countertop.

**AdaIAT**: The image shows a bunch of bananas placed on a round, light-colored plate. The bananas appear ripe with some brown spots visible on their yellow peels, indicating they are ready to eat. In the background, there is a kitchen countertop with various items on it, including a glass container that appears to be a spice jar and a small white object that could be a spoon or a kitchen utensil. The lighting in the room is warm, giving the scene a cozy feel.

Please describe the image in detail.

**Greedy:** The image is a black-and-white photograph of a partially eaten sandwich on a white plate. The sandwich appears to be made with a toasted bun, and its filling includes what looks like a meat patty or cutlet, possibly with some sauce or condiments visible. The bread has a slightly charred exterior, suggesting it was grilled or toasted. In the background, there's a blurred view of a person, but their features are not discernible due to the focus being on the sandwich. To the right side of the frame, there is a small white bowl, likely containing a dipping sauce or condiment for the sandwich. The overall composition gives a casual dining atmosphere, and the monochrome effect adds a timeless quality to the image.

**IAT**: The image is a black-and-white photograph that appears to be taken in a casual dining setting, possibly a café or restaurant. The main focus of the photo is a partially eaten sandwich placed on a white plate. The sandwich has a dark bread roll and seems to contain some kind of filling, which could include meat, cheese, or vegetables. In the background, there's a blurred view of what looks like a person sitting at a table, though their features are not clear due to the depth of field and the monochrome filter. The overall atmosphere suggests a relaxed mealtime with soft lighting, contributing to the cozy ambiance of the scene.

**AdaIAT**: The image is a black-and-white photograph of a partially eaten sandwich on a white plate. The sandwich appears to be made with a soft, possibly toasted bread, and it seems to contain some kind of filling that looks like a mixture of ingredients, possibly including cheese or a creamy substance. The sandwich is cut in half, and one half is missing a bite, indicating that someone has already started eating it.\n\nIn the background, there's a blurred setting that suggests an indoor environment, such as a restaurant or café. The focus of the image is clearly on the sandwich, which is placed centrally on the plate. The lighting in the photo creates a high-contrast effect, emphasizing the textures of the bread and the filling. The overall composition gives a casual and appetizing feel to the meal.

Figure 14. Qualitative results of hallucination mitigation in Qwen2.5-VL-7B, with hallucinated segments highlighted in **red**.

## References


[1] W. An, F. Tian, S. Leng, et al. Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29915–29926, 2025. 1

[2] S. Bai, K. Chen, X. Liu, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1

[3] X. Chen, Z. Wu, X. Liu, et al. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025. 1

[4] Q. Huang, X. Dong, P. Zhang, et al. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427, 2024. 1

[5] Z. Jiang, J. Chen, B. Zhu, et al. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25004–25014, 2025. 1

[6] S. Leng, H. Zhang, G. Chen, et al. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882, 2024. 1

[7] T. Lin, M. Maire, S. Belongie, et al. Microsoft coco: Common objects in context. In *Proceedings of the European conference on computer vision*, pages 740–755. Springer, 2014. 1

[8] H. Liu, C. Li, Y. Li, et al. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 1

[9] S. Liu, K. Zheng, and W. Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *Proceedings of the European Conference on Computer Vision*, pages 125–140. Springer, 2024. 1

[10] D. Podell, Z. English, K. Lacey, et al. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2

[11] H. Touvron, L. Martin, K. Stone, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 2

[12] Y. Zhou, C. Cui, J. Yoon, et al. Analyzing and mitigating object hallucination in large vision-language models. In *Proceedings of theThe Twelfth International Conference on Learning Representations*, 2024. 1