

# Residual Decoding: Mitigating Hallucinations in Large Vision-Language Models via History-Aware Residual Guidance

## Supplementary Material

### A. Implementation Details

#### A.1. Datasets

**MME [16]** evaluates the capabilities of MLLMs, dividing the evaluation into two major categories: perception and cognition. The perception category includes fine-grained tasks such as existence, count, location, rough color, poster, celebrity, scene, landmark, artwork identification, and OCR. The cognition category includes tasks like commonsense reasoning, numerical calculations, text translation, and code reasoning. All questions in this benchmark are structured to be answered with a simple yes or no.

**POPE [32]** is a metric based on Visual Question Answering (VQA) designed to evaluate hallucinations in MLLMs. This metric assesses the model’s response to the question, “Is [object] in this image?” To clarify that this is a binary VQA task, the prompt is supplemented with the instruction, “Please answer yes or no.” POPE is built on three core data sources: COCO, A-OKVQA, and GQA. We report the overall accuracy and F1-score.

**CHAIR [51]** is used to assess the alignment between generated captions and the content of the given image. The CHAIR metric consists of two evaluation types: per-instance evaluation ( $\text{CHAIR}_I$ ) and per-sentence evaluation ( $\text{CHAIR}_S$ ), which are defined as follows:

$$\text{CHAIR}_I = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all objects mentioned}\}|} \quad (11)$$

$$\text{CHAIR}_S = \frac{|\{\text{sentences with hallucinated object}\}|}{|\{\text{all sentences}\}|} \quad (12)$$

**HallusionBench [20].** HallusionBench is a benchmark constructed to evaluate hallucination robustness of large vision–language models under challenging, misleading scenarios. It contains 951 carefully curated question–answer pairs, where each question is designed to probe whether a model can resist visually or semantically deceptive cues and avoid generating hallucinated content. This benchmark thus provides a focused testbed for analyzing and improving the reliability of multimodal reasoning.

**MMBench [45]** is a comprehensive benchmark for evaluating general purpose vision language models. It contains

carefully curated multiple choice questions in both English and Chinese that test perception, reasoning, and understanding across diverse real world images and scenarios.

**ScienceQA [46]** is a multimodal multiple-choice benchmark that evaluates scientific reasoning over grade-school curricula, with questions accompanied by textual context and often images or diagrams. In our experiments, we adopt the official ScienceQA test split, which consists of 2,017 question–answer pairs, and use it to assess the model’s ability to perform multimodal reasoning and science knowledge understanding.

**MMStar [8]** is a comprehensive benchmark for evaluating large vision–language models, consisting of 1,500 carefully curated multiple-choice questions that cover a wide range of real-world images and tasks. It is designed to probe fine-grained multimodal abilities, including perception, reasoning, and instruction following, thereby enabling more reliable and nuanced assessment of general-purpose multimodal models.

**SEEDBench2\_Plus [30]** is a large scale benchmark for evaluating multimodal large language models. It contains 2,277 question–answer pairs covering diverse perception and reasoning skills, and provides a fine grained assessment of both visual understanding and high level multimodal reasoning.

**MM-Vet [66]** is a benchmark designed to assess large vision–language models on challenging, multi-step tasks. It specifies six core capabilities in the vision–language setting: recognition, optical character recognition, factual and commonsense knowledge, language generation, spatial understanding, and mathematical reasoning. To obtain consistent and reliable scores across heterogeneous question formats, it employs a language model as an automatic evaluator. The benchmark contains 187 images collected from diverse online sources and 205 questions, each constructed so that answering it requires one or more of the defined capabilities.

**MMVP.** MMVP [60] is a benchmark targeting the fine-grained visual recognition ability of LVLMs based on CLIP-blind image pairs. It contains 150 pairs of images, with each pair accompanied by a binary-choice question.

The two images in a pair are evaluated independently, and an LVLMM receives credit for a pair only if it answers the questions for both images correctly.

**LLaVA-Bench.** LLaVA-Bench (In-the-Wild) [42] consists of 24 images representing various complex scenes, memes, paintings, and sketches, paired with 60 challenging questions. A subset of this dataset is used for qualitative comparisons of responses generated by different decoding strategies. We assess the accuracy and depth of the generated responses using GPT-4V.

## B. Theoretical Analysis of ResDec

### B.1. Derivation of Language Priors

Following Lin et al. [39], we define the *language prior* as the text-only, vision-agnostic conditional distribution  $P(y | x)$ , which represents the generative tendency of the output sequence  $y = (y_1, \dots, y_T)$  given only the textual context  $x$ , without considering the visual input  $v$ . This language prior is related to the joint distribution of text and vision via the equivalent factorizations:

$$P(y, v | x) = P(y | x) P(v | y, x) = P(v | x) P(y | v, x), \quad (13)$$

where  $P(y | x)$  isolates the language-driven preference over the outputs once the visual evidence  $v$  is removed. These factorizations highlight how the text-only distribution and visual context interact to form the complete generative process. At the token level, both the visual-conditional likelihood  $P(y | v, x)$  and the language prior  $P(y | x)$  admit autoregressive factorizations, meaning that the probability of the entire sequence is factorized into the product of the probabilities of each individual token conditioned on the previous ones:

$$\begin{aligned} P(y | v, x) &= \prod_{t=1}^T P(y_t | y_{<t}, v, x), \\ P(y | x) &= \prod_{t=1}^T P(y_t | y_{<t}, x), \end{aligned} \quad (14)$$

The first equation represents the likelihood of the output sequence given both the visual and textual context, while the second equation shows the likelihood given only the textual context. These factorizations treat each token in the sequence independently, conditioned on all previous tokens, allowing for sequential prediction. To compare the visual-conditional likelihood and the language prior, we combine the joint factorizations with Bayes' rule. This results in a ratio that isolates the effect of the visual input relative to the language prior:

Models	Historical Window $W$	POPE		MME
		Accuracy $\uparrow$	F1 $\uparrow$	Accuracy $\uparrow$
LLaVA-1.5	2	87.05	86.59	1869.23
	4	84.45	82.71	1762.76
	8	87.23	86.93	1881.57
	16	86.96	86.32	1882.41
	32	87.35	87.03	1880.86
Qwen2.5-VL	2	84.33	85.21	2232.54
	4	86.86	85.63	2326.93
	8	90.16	89.56	2348.40
	16	90.33	89.96	2350.52
	32	89.98	89.73	2346.81

Table 7. Performance of predefined historical window  $W$  on LLaVA-1.5/Qwen2.5-VL for POPE and MME benchmarks.

$$\frac{P(y | v, x)}{P(y | x)} = \frac{P(y, v | x)}{P(y | x) P(v | x)} = \frac{P(v | y, x)}{P(v | x)}, \quad (15)$$

This ratio quantifies how much the visual input  $v$  contributes to the generation process beyond the language prior  $P(y | x)$ . The ratio serves as a *PMI-style* measure of the influence of the visual context. Taking the logarithm of this ratio turns it into a token-wise subtraction between the visual-conditional and the text-only next-token scores at each timestep  $t$ :

$$\log \frac{P(y | v, x)}{P(y | x)} = \sum_{t=1}^T \left[ \log P(y_t | y_{<t}, v, x) - \log P(y_t | y_{<t}, x) \right], \quad (16)$$

This formulation allows us to evaluate how much the visual context adjusts the prediction for each token, removing the contribution of the language prior. Finally, the language prior itself can be expressed as a marginalization over the visual variable  $v$ , which accounts for all possible visual contexts. This marginalization is given by:

$$P(y | x) = \mathbb{E}_{v \sim P(v|x)} [P(y | v, x)], \quad (17)$$

where the expectation is taken over all possible visual inputs  $v$ , conditioned on the textual context  $x$ . In practice, this expectation is approximated using Monte Carlo sampling, which gives:

$$P(y | x) \approx \frac{1}{n} \sum_{j=1}^n P(y | v_j, x), \quad (18)$$

where  $\{v_j\}_{j=1}^n$  are samples drawn from the image distribution conditioned on  $x$ . This Monte Carlo approximation effectively captures the language prior while accounting for the variability in the visual context.

## B.2. Theoretical Analysis

**Definitions.** Let  $V$  be the visual variable,  $H$  the history (image tokens + past text), and  $Y$  the next-token output. Denote the baseline decoder by  $p(y | v, h)$  and the residual-induced distribution by  $r(y | v, h)$ . ResDec forms the  $\alpha$ -blended channel

$$p_\alpha(y | v, h) := \frac{p(y | v, h)^{1-\alpha} r(y | v, h)^\alpha}{Z_\alpha(v, h)},$$

$$Z_\alpha(v, h) := \sum_{y'} p(y' | v, h)^{1-\alpha} r(y' | v, h)^\alpha.$$

Write the  $H$ -marginal as

$$p_\alpha(y | h) := \sum_v p_\alpha(y | v, h) p(v | h).$$

The conditional mutual information is

$$I_\alpha(V; Y | H) := \mathbb{E}_h [\text{KL}(p_\alpha(\cdot | v, h) \| p_\alpha(\cdot | h))].$$

Define the *residual advantage*

$$A(y, v, h) := \log r(y | v, h) - \log r(y | h).$$

**Theorem 1.** Assume regularity so that differentiation under the expectation is valid, and suppose the residual is informative in the sense that for some  $\varepsilon > 0$ ,

$$\mathbb{E}_h \mathbb{E}_{v|h} \mathbb{E}_{y \sim p(\cdot | v, h)} [A(y, v, h)] \geq \varepsilon.$$

Then the directional derivative of the conditional MI at  $\alpha = 0$  satisfies

$$\left. \frac{d}{d\alpha} I_\alpha(V; Y | H) \right|_{\alpha=0} = \mathbb{E}_h \mathbb{E}_{v|h} \mathbb{E}_{y \sim p(\cdot | v, h)} [A(y, v, h)] \geq \varepsilon.$$

Consequently, there exists  $\alpha_0 > 0$  such that for all  $\alpha \in (0, \alpha_0]$ ,

$$I_\alpha(V; Y | H) \geq I_0(V; Y | H) + \frac{\varepsilon}{2} \alpha.$$

**Proof.** Rewrite  $p_\alpha$  as an exponential tilt of  $p$ :

$$p_\alpha(y | v, h) = \frac{p(y | v, h) \exp(\alpha U_v(y))}{Z_\alpha(v, h)},$$

$$U_v(y) := \log r(y | v, h) - \log p(y | v, h).$$

Similarly,

$$p_\alpha(y | h) = \frac{p(y | h) \exp(\alpha \bar{U}(y))}{\bar{Z}_\alpha(h)},$$

$$\bar{U}(y) := \log r(y | h) - \log p(y | h).$$

For the log-densities, exponential-family calculus gives

$$\left. \frac{\partial}{\partial \alpha} \log p_\alpha(y | v, h) \right|_0 = U_v(y) - \mathbb{E}_{y' \sim p(\cdot | v, h)} [U_v(y')],$$

$$\left. \frac{\partial}{\partial \alpha} \log p_\alpha(y | h) \right|_0 = \bar{U}(y) - \mathbb{E}_{y' \sim p(\cdot | h)} [\bar{U}(y')].$$

Differentiating the KL form of  $I_\alpha$  at  $\alpha = 0$  and using  $\frac{d}{d\alpha} \text{KL}(P_\alpha \| Q_\alpha) = \mathbb{E}_{P_\alpha} [\partial_\alpha \log P_\alpha - \partial_\alpha \log Q_\alpha]$  at  $\alpha = 0$  yields

$$\left. \frac{d}{d\alpha} I_\alpha \right|_0 = \mathbb{E}_h \mathbb{E}_{v|h} \mathbb{E}_{y \sim p(\cdot | v, h)} [(U_v(y) - \bar{U}(y))].$$

With the definitions of  $U_v$  and  $\bar{U}$ , the bracket equals

$$\begin{aligned} & (\log r(y | v, h) - \log p(y | v, h)) \\ & - (\log r(y | h) - \log p(y | h)) \\ & = A(y, v, h) - \text{PMI}_p(y; v | h). \end{aligned}$$

Taking expectation over  $y \sim p(\cdot | v, h)$  cancels the  $\text{PMI}_p$  term, giving the stated derivative. The claimed local increase follows by continuity.

**Theorem 2.** Let  $v$  be the visual input,  $x$  the text query, and  $h$  the decoding history. Denote the base logits by  $\ell(y | v, x, h)$  and the residual offsets by  $r_h(y)$ . Define the base and residual-augmented conditional distributions

$$p_0(y | v, x, h) = \text{softmax}(\ell(y | v, x, h)),$$

$$p_\alpha(y | v, x, h) = \text{softmax}(\ell(y | v, x, h) + \alpha r_h(y)),$$

with  $\alpha \in (0, 1]$ . Equivalently,

$$p_\alpha(y | v, x, h) = \frac{p_0(y | v, x, h) \exp(\alpha r_h(y))}{Z(\alpha)},$$

$$Z(\alpha) = \sum_{y'} p_0(y' | v, x, h) \exp(\alpha r_h(y')).$$

Assume the residual aligns with the desirable visual grounding in the sense that  $\text{Cov}_{p_\alpha}(r_h(Y), \log p_0(Y | v, x, h)) \geq 0$  for  $\alpha \in [0, 1]$ , and  $\text{Var}_{p_\alpha}(r_h(Y)) > 0$ . Then, for any  $\alpha \in (0, 1]$ , the conditional Shannon entropy strictly decreases:

$$H(Y | v, x, h)_{p_\alpha} < H(Y | v, x, h)_{p_0}.$$

**Proof.** Use the exponential-tilted form of  $p_\alpha$  displayed above and write the conditional entropy under  $p_\alpha$  as

$$\begin{aligned} H_\alpha &= - \sum_y p_\alpha(y) \log p_\alpha(y) \\ &= - \mathbb{E}_{p_\alpha} [\log p_0(Y)] - \alpha \mathbb{E}_{p_\alpha} [r_h(Y)] + \log Z(\alpha). \end{aligned}$$

Differentiate with respect to  $\alpha$ . Invoking the standard score function identity for the  $\alpha$ -indexed family  $\{p_\alpha\}$ , namely  $\frac{d}{d\alpha} \mathbb{E}_{Y \sim p_\alpha} [f(Y)] = \text{Cov}_{Y \sim p_\alpha}(f(Y), r_h(Y))$  for any integrable test function  $f$ , and using  $\frac{d}{d\alpha} \log Z(\alpha) = \mathbb{E}_{Y \sim p_\alpha} [r_h(Y)]$ , we obtain

$$\begin{aligned} \frac{d}{d\alpha} H_\alpha &= - \text{Cov}_{p_\alpha}(\log p_0(Y), r_h(Y)) - \mathbb{E}_{p_\alpha} [r_h(Y)] \\ &\quad - \alpha \text{Cov}_{p_\alpha}(r_h(Y), r_h(Y)) + \mathbb{E}_{p_\alpha} [r_h(Y)]. \end{aligned}$$

The middle expectations cancel, yielding

$$\frac{d}{d\alpha} H_\alpha = -\text{Cov}_{p_\alpha}(\log p_0(Y), r_h(Y)) - \alpha \text{Var}_{p_\alpha}(r_h(Y)).$$

Under the alignment assumption,  $r_h(Y)$  is positively associated with the base log-likelihood  $\log p_0(Y)$ , hence  $\text{Cov}_{p_\alpha}(\log p_0(Y), r_h(Y)) \geq 0$ . Moreover, the residual signal is non-degenerate under  $p_\alpha$ , so  $\text{Var}_{p_\alpha}(r_h(Y)) > 0$ . These conditions hold for all  $\alpha$  and together capture positive association with nontrivial variability; therefore, for any  $\alpha > 0$ ,

$$\frac{d}{d\alpha} H_\alpha < 0.$$

Integrating from 0 to any  $\alpha \in (0, 1]$  gives

$$H_\alpha - H_0 = \int_0^\alpha \frac{d}{ds} H_s ds < 0,$$

which is precisely the claimed entropy decrease.

## C. Additional Ablation Studies

### C.1. Predefined Historical Window

In this section, we conduct experiments on the Predefined Historical Window  $W$ . Since  $W$  merely serves as a parameter for identifying the U-shape, we set  $W$  to 2, 4, 8, 16, and 32. The corresponding experimental results are presented in Table 7. We find that the U-shape is typically achievable when  $W = 8$ ; thus, we usually set  $W = 8$  in our experiments.

### C.2. Historical Window Selection Strategies

As discussed in Sec. 3.2, the temporal evolution of candidate tokens can be decomposed into three phases: the Pre-Semantic Clarity Phase (PSAP), the Semantic Anchoring Phase (SAP), and the Expressive Divergence Phase (EDP). To validate the necessity of leveraging both SAP and EDP for historical information aggregation, we ablate the window selection on Qwen2.5-VL and LLaVA-1.5 using the POPE and MME benchmarks.

We compare our *ResDec* strategy against using the Full Window (PSAP+SAP+EDP), individual phases (PSAP only, SAP only, EDP only), and a heuristic Top-4 Confident Steps selection. As shown in Table 8, incorporating the PSAP degrades performance. This degradation is likely due to the candidate token distribution transitioning from an initial state of disorder, which introduces early noise before settling on core semantics. While relying solely on the SAP improves accuracy, it slightly trails the full *ResDec* strategy. This confirms our observation that incorporating the EDP—which contains effective decoding guidance due to its diverse expressions—provides vital contextual information for maintaining coherent generation.

Models	Strategy	POPE		MME
		ACC ↑	F1 ↑	ACC ↑
Qwen2.5-VL	Full Window (PSAP+SAP+EDP)	85.83	84.72	2275.68
	PSAP Only	85.17	84.24	2151.31
	SAP Only	88.82	87.96	2337.68
	EDP Only	86.14	84.27	2330.47
	Top-4 Confident Steps <i>ResDec</i> (ours)	<b>90.16</b>	<b>89.56</b>	<b>2348.40</b>
LLaVA-1.5	Full Window (PSAP+SAP+EDP)	84.87	84.03	1813.62
	PSAP Only	84.13	83.32	1797.94
	SAP Only	86.21	85.87	1846.45
	EDP Only	85.56	84.74	1843.32
	Top-4 Confident Steps <i>ResDec</i> (ours)	<b>87.23</b>	<b>86.93</b>	<b>1881.57</b>

Table 8. Ablation study on historical window selection strategies.

### C.3. Aggregation Strategy

*ResDec* employs a confidence-weighted aggregation method based on a local confidence metric derived from the model’s internal token distribution. To justify this proposed pooling strategy, we compare it against two alternative heuristic aggregation methods: Uniform (Mean) Pooling and Distance Decay (inverse temporal-distance weighting).

The results presented in Table 9 demonstrate that *ResDec*’s confidence-weighted pooling consistently outperforms both uniform and distance-based decay strategies. By dynamically weighting historical time steps based on their degree of certainty, *ResDec* more effectively isolates stable semantics and mitigates language-prior hallucinations compared to fixed or strictly distance-based heuristic weightings.

Models	Aggregation Strategy	POPE		MME
		ACC ↑	F1 ↑	ACC ↑
Qwen2.5-VL	Uniform (Mean) Pooling	88.31	87.21	2342.76
	Distance Decay	86.93	85.45	1954.79
	<i>ResDec</i> (ours)	<b>90.16</b>	<b>89.56</b>	<b>2348.40</b>
LLaVA-1.5	Uniform (Mean) Pooling	85.26	84.74	1785.41
	Distance Decay	83.21	82.55	1731.62
	<i>ResDec</i> (ours)	<b>87.23</b>	<b>86.93</b>	<b>1881.57</b>

Table 9. Ablation study on the aggregation strategy.

## D. More Case Studies

Additional case studies are presented to illustrate the effectiveness of our *ResDec*.

[Instruction] Please describe this image in detail.

[Regular Decoding] The image features a green park bench with a bottle of water placed on it. The bench is located near a sidewalk, and there are several people walking by. In the background, a car can be seen parked on the street. There are three people in the scene, with one person walking closer to the left side of the bench, another person walking further to the right, and the third person walking behind the bench. Additionally, there is a **handbag placed on the ground** near the person walking on the left side of the bench.



[Residual Decoding] The image features a green bench in a park, with a blue bottle sitting on the bench, close to the back and the right side. The bench is positioned near a sidewalk, and a street with parked cars can be seen in the background. There are a few people in the scene, with one person standing close to the left side of the bench and, in the background, two more people, one near the left and the other near the right side of the park.

[Instruction] Please describe this image in detail.

[Regular Decoding] The image features a large, beautifully decorated cake with white frosting and red lettering. The cake is placed on a dining table, and it is adorned with a knife and a **fork**, ready to be cut and served. The cake is **shaped like a heart** and is likely meant to celebrate a special occasion or event. The red lettering on the cake reads "welcome malachi," indicating that it is a welcoming gesture for someone named Malachi.



[Residual Decoding] The image showcases a large cake decorated beautifully, with a red frosting and white frosting to celebrate a welcoming. The cake is topped with a red cake knife, and a decorative cake topper, a red cake topper with a cake decor to decorate the cake. The cake is placed on a dining table, and a cake serving spatula is also visible on the table, ready to be used to cut and serve the cake.

[Instruction] Please describe this image in detail.

[Regular Decoding] The image features a cozy dining room with a dining table set for a meal. The table is adorned with several wine glasses, cups, and a **bottle of wine**. There are also a couple of forks and knives placed on the table. The dining area is furnished with a couch and a chair, providing a comfortable atmosphere for guests to enjoy their meal. The room is decorated with a fireplace, adding warmth and ambiance to the space. **A clock is mounted on the wall, and a vase can be seen on the table**, adding to the room's decor. The dining table is surrounded by chairs, and a couch is situated nearby, creating a welcoming and inviting environment for guests to relax and dine.



[Residual Decoding] The image showcases a beautifully decorated dining room, furnished with a dining table and a set of chairs. The table is set with dishes, plates, and wine glasses, ready for a meal. The dining table, with a tablecloth, and chairs surround it, and the setting is inviting and elegant. There are several wine glasses and cups on the table, and a bottle, possibly a wine bottle, is placed on the table as well. The dining table, with a well-arranged set, and the chairs set for a meal, create a cozy and inviting atmosphere for a dining experience.

[Instruction] Please describe this image in detail.

[Regular Decoding] The image features a man standing on the back of a boat, painting a crab on the side of the vessel. The boat is white and blue, and it appears to be a fishing boat. There are several other people on the boat, **some of them standing near the man painting the crab, while others are scattered around the boat**. In addition to the people on the boat, there are a few objects visible in the scene. **A bottle can be seen on the boat, and a cup is placed nearby**. The presence of these items suggests that the people on the boat might be enjoying a leisurely time together.



[Residual Decoding] The image depicts a man a boat painted with a crab on the side, a crab painted on the boat. The man is standing on the boat, painting the crab on the side, while a crab painted on the boat nearby. There are several other people on the boat, and a smaller boat can be seen in the background. The boat, with the crab painted on it, is the main focus of the scene, and the man painting the crab on the side is the most prominent figure in the image.

[Instruction] Please describe this image in detail.

[Regular Decoding] The image features a small bathroom with a white toilet and a urinal mounted on the wall. The toilet is positioned on the left side of the bathroom, while the urinal is located on the right side. **The toilet seat is up, revealing the toilet bowl. There are two toilet paper holders in the bathroom, one on the left side and the other on the right side. The toilet paper on the left side is placed on the holder, while the toilet paper on the right side is hanging from the holder.** The bathroom appears to be clean and well-maintained.



[Residual Decoding] The image showcases a bathroom stall with a toilet and a urinal, both white in color. The toilet and urinal are positioned close to each other, with the urinal to the right of the toilet. A toilet paper dispenser is mounted on the wall above the toilet, and a toilet paper roll is hanging on the dispenser. The toilet paper dispenser is also next to the urinal to the right of the toilet. In addition to the toilet and urinal, the bathroom stall features a sink, and a toilet paper dispenser is also mounted on the wall above the sink.

[Instruction] Please describe this image in detail.

[Regular Decoding] The image features a man shearing a sheep **in a barn**. The man is standing next to the sheep, which is lying on the ground, and he is using a pair of shears to remove the wool. The sheep appears to be a large, fluffy animal, and the man is focused on his task. In the background, **there are two other sheep, one on the left side and another on the right side of the image. Additionally, there are two people present in the scene, one standing near the left side of the image and another person standing near the right side. A bottle can also be seen on the ground, possibly containing a shearing tool or a cleaning solution.**



[Residual Decoding] The image depicts a man shearing a sheep in a sheep shearing shed. The sheep is on the ground, and the man is shearing the wool with shears in his hands. The shearing process is in progress, and the man is focused on the task. There are several other sheep in the background, and a nearby fence can be seen, enclosing the area. The man and the sheep in the foreground are the main focus of the scene, showcasing the shearing process in action.

Figure 6. More examples comparing regular decoding and residual decoding in image captioning. Hallucinated details in “regular” decoding outputs are highlighted in red.



[Question] Is there a baseball bat in the image?

LLaVA-1.5 [Regular Decoding] Yes.

LLaVA-1.5 [Residual Decoding] No.



[Question] Is there a bicycle in the image?

LLaVA-1.5 [Regular Decoding] No.

LLaVA-1.5 [Residual Decoding] Yes.



[Question] Is there a wine glass in the image?

LLaVA-1.5 [Regular Decoding] Yes.

LLaVA-1.5 [Residual Decoding] No.



[Question] Is there a bottle in the image?

LLaVA-1.5 [Regular Decoding] No.

LLaVA-1.5 [Residual Decoding] Yes.



[Question] Is there a stop sign in the image?

LLaVA-1.5 [Regular Decoding] Yes.

LLaVA-1.5 [Residual Decoding] No.

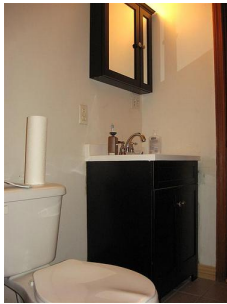


[Question] Is there a stop sign in the image?

LLaVA-1.5 [Regular Decoding] Yes.

LLaVA-1.5 [Residual Decoding] No.

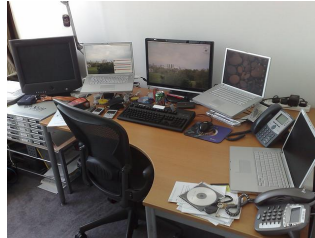
Figure 7. More examples comparing regular decoding and residual decoding in image captioning. Hallucinated details in “regular” decoding outputs are highlighted in red.



[Question] Is there a toothbrush in the image?

Qwen2.5-VL [Regular Decoding] No.

Qwen2.5-VL [Residual Decoding] Yes.



[Question] Is there a mouse in the image?

Qwen2.5-VL [Regular Decoding] No.

Qwen2.5-VL [Residual Decoding] Yes.



[Question] Is there a bowl in the image?

Qwen2.5-VL [Regular Decoding] Yes.

Qwen2.5-VL [Residual Decoding] No.



[Question] Is there a truck in the image?

Qwen2.5-VL [Regular Decoding] No.

Qwen2.5-VL [Residual Decoding] Yes.



[Question] Is there a backpack in the image?

Qwen2.5-VL [Regular Decoding] No.

Qwen2.5-VL [Residual Decoding] Yes.



[Question] Is there a bench in the image?

Qwen2.5-VL [Regular Decoding] Yes.

Qwen2.5-VL [Residual Decoding] No.

Figure 8. More examples comparing regular decoding and residual decoding in image captioning. Hallucinated details in “regular” decoding outputs are highlighted in red.