Where MLLMs Attend and What They Rely On: Explaining Autoregressive Token Generation

Supplementary Material

6. EAGLE Algorithm

The detailed calculation process of the proposed EAGLE algorithm is outlined below.

Algorithm 1: EAGLE: Explaining Autoregressive Generation by Language priors or Evidence in multimodal large language models (MLLMs)

Input: Image $\mathbf{I} \in \mathbb{R}^{h \times w \times 3}$, partitioning algorithm $\texttt{Div}(\cdot)$, prompt $\texttt{Prompt}$, generated sequence $\mathbf{y}$, target token positions T, vocabulary indices $\mathcal{V}$.

Output: Ordered subset π, saliency map $\mathcal{A} \in \mathbb{R}^{h \times w}$, influence scores I_t.

$V \leftarrow \texttt{Div}(\mathbf{I})$;
$\pi \leftarrow \varnothing$; /* Initialize ordered subset */
$\mathcal{A}_1 \leftarrow 0$;
for $i = 1$ **to** $|V|$ **do**
  $S_d \leftarrow V \setminus S$;
  $\alpha \leftarrow \arg\max_{\alpha \in S_d} \mathcal{F}(\pi \cup \{\alpha\})$;
  $\pi \leftarrow \pi \,\|\, \{\alpha\}$;
  if $i > 1$ **then**
    $\mathcal{A}_i \leftarrow \mathcal{A}_{i-1} - \left|\mathcal{F}(\pi_{:i}) - \mathcal{F}(\pi_{:i-1})\right|$; /* Saliency update */
end
for $i = 1$ **to** $|T|$ **do**
  $s_{\max} \leftarrow \max_{1 \le j \le |\pi|} p(y_{t_i} = v_i \mid \pi_{:j}, \texttt{Prompt}, \mathbf{y}_{<t_i})$;
  $I_{t_i} \leftarrow \sum_{r=1}^{|\pi|} \Big(s_{\max} - p(y_{t_i} = v_i \mid \pi_{:r}, \texttt{Prompt}, \mathbf{y}_{<t_i}) \Big)$; /* Language prior vs. perception evidence */
end
return π, norm($\mathcal{A}$), norm(I_t)

7. Additional Experimental Details

For the image captioning task on MS COCO, the prompt used for all MLLMs is:

```
Describe the image in one factual English sentence of no more than 20 words. Do not include information that is not clearly visible.
```

For the hallucination detection task on RePOPE, the prompt used is:

```
You are asked a visual question answering task.
First, answer strictly with "Yes" or "No".
Then, provide a short explanation if necessary.

Question: {question}
Answer:
```

8. Limitations and Future Works.

Despite its effectiveness, our work has two main limitations. First, the iterative subset selection and greedy search limit scalability compared with lightweight visualization methods. Nevertheless, our approach provides a promising interpretable pathway and clarifies a potential upper bound of attribution for MLLMs; in future work, we will design more efficient attribution algorithms. Second, the framework focuses on hallucination explanation and partial mitigation, leaving proactive prevention unexplored. In future work, we will leverage explanations to automatically detect and mitigate hallucinations and, once failure modes are identified, develop data-/parameter-efficient methods for minimal-cost model repair.

9. Additional Qualitative Results

In this appendix, we provide extended qualitative visualizations that complement the main findings in Fig. 3, Fig. 4, and Fig. 5. These supplementary results aim to offer a finer-grained perspective on how competing attribution methods and our proposed approach behave across diverse settings. Specifically, we present: (i) sentence-level explanations on both MS COCO and MMVP, (ii) word-level explanations on MS COCO, and (iii) hallucination attribution visualizations on additional samples. Collectively, these results provide deeper insights into the consistency, precision, and interpretability of our method.

9.1. Sentence-level Explanations on MS COCO and MMVP

As shown in Fig. 6 and Fig. 7, our method produces faithful explanations for **LLaVA-1.5** by tightly aligning highlighted regions with relevant caption tokens (e.g., "smiling," "hat," "motor") or VQA queries (e.g., "Is the shark's belly visible?"). In contrast, LLaVA-CAM often distributes attention diffusely across the scene, while IGOS++ over-activates irrelevant background regions.

For **Qwen2.5-VL**, Fig. 8 and Fig. 9 show that our method generates concise and semantically meaningful attribution maps. For example, in captions mentioning multiple objects, our approach selectively highlights the relevant ones while avoiding redundancy. In VQA tasks, it accurately isolates queried entities such as a remote button, whereas baselines either miss the target or introduce noise.

Similarly, for **InternVL3.5** (Fig. 10, Fig. 11), our method highlights precise object-centric regions corresponding to key caption tokens (e.g., "sandwich," "frisbee") and VQA queries (e.g., "Does the snowman have arms made of branches?"). Baseline methods either scatter attention broadly or fail to capture the queried object, reducing interpretability. These results collectively demonstrate that our approach consistently improves faithfulness and transparency across different models and datasets.

9.2. Word-level Explanations on MS COCO

Beyond sentence-level results, we further evaluate our method at the word-level with ground-truth bounding boxes. Fig. 12, Fig. 13, and Fig. 14 illustrate that our method produces sparse yet highly accurate localization of queried objects such as "boat," "keyboard," or "truck." By contrast, IGOS++ frequently covers overly broad regions, while LLaVA-CAM and TAM often fail to precisely localize objects. These comparisons highlight the advantage of our method in generating interpretable, object-centric attributions.

9.3. Additional Hallucination Attribution Visualizations

We also provide supplementary hallucination attribution results on MS COCO (Fig. 15, Fig. 16, Fig. 17). Unlike the main paper, these figures focus exclusively on our method to illustrate how it identifies hallucination-prone regions across diverse queries.

For **LLaVA-1.5** (Fig. 15), hallucinations typically arise from visually similar structures. For example, queries about a "snowboard" lead to confusions with surfboard-like regions, while small background cues induce false detections for "traffic light" or "cup." Our attribution maps isolate these exact regions, providing interpretable evidence of failure modes.

For **Qwen2.5-VL** (Fig. 16), hallucinations are often caused by small or occluded objects. For instance, reflective regions resembling a phone screen mislead the model when asked about "cell phones," while circular patterns in the background induce false positives for "bicycle." Our approach sharply localizes these misleading cues, enhancing transparency.

Finally, for **InternVL3.5** (Fig. 17), hallucinations are triggered by overlapping or occluded objects. For example, confusion between a fork and a spoon is precisely localized, as are reflective regions falsely identified as "TVs" or cluttered areas misinterpreted as "dining tables." These examples underscore the effectiveness of our method in diagnosing hallucination sources in a fine-grained and transparent manner.

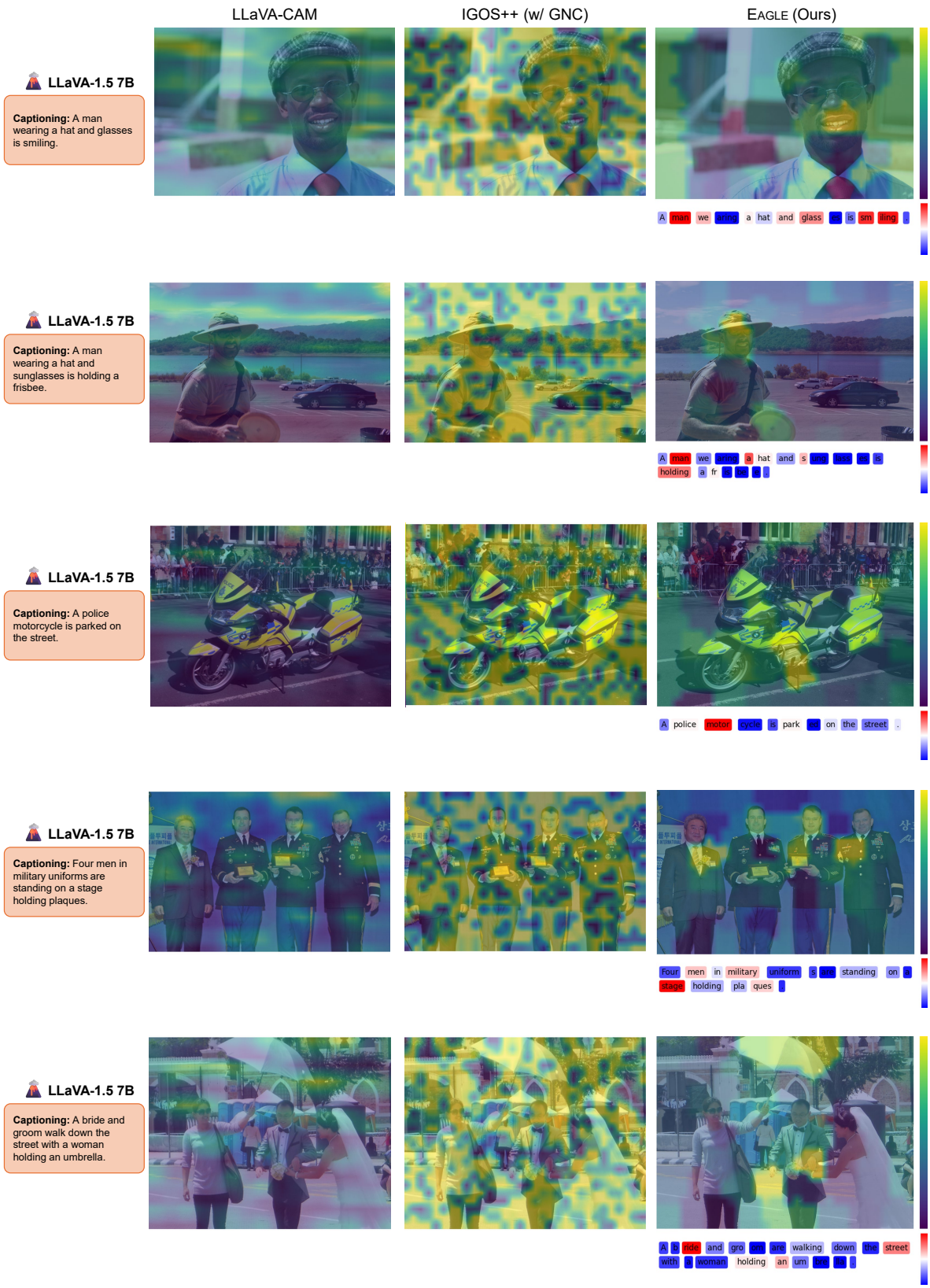

Figure 6. Sentence-level explanation results for **LLaVA-1.5** on the MS COCO dataset. Our method consistently identifies semantically critical regions that align with highlighted tokens in the caption, while baseline methods either fail to capture relevant areas (LLaVA-CAM) or over-highlight irrelevant background regions (IGOS++).

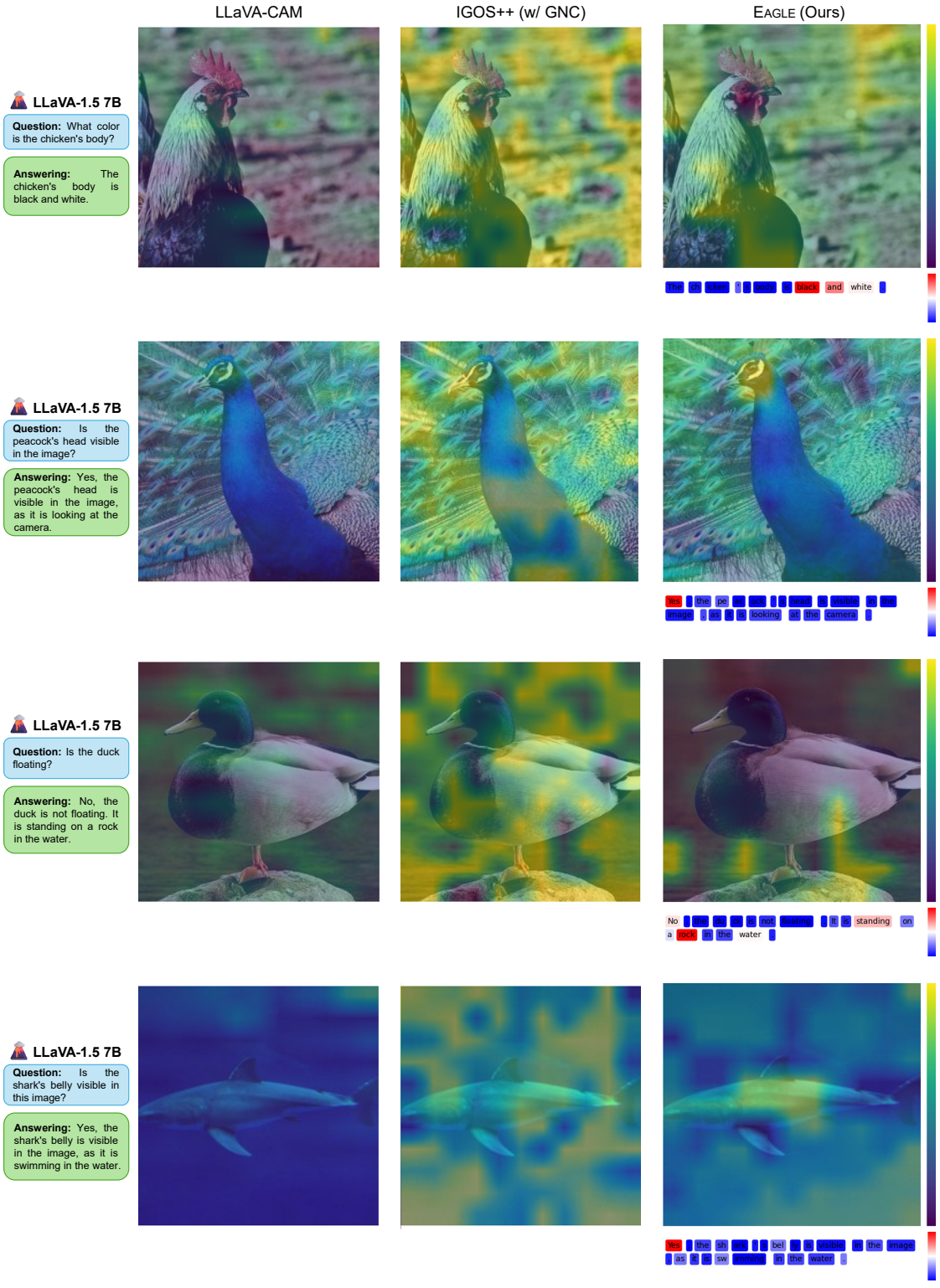

Figure 7. Sentence-level explanation results for **LLaVA-1.5** on the MMVP dataset. Compared to the baselines, our method highlights regions that are directly related to the VQA queries, resulting in explanations that are more interpretable and trustworthy.

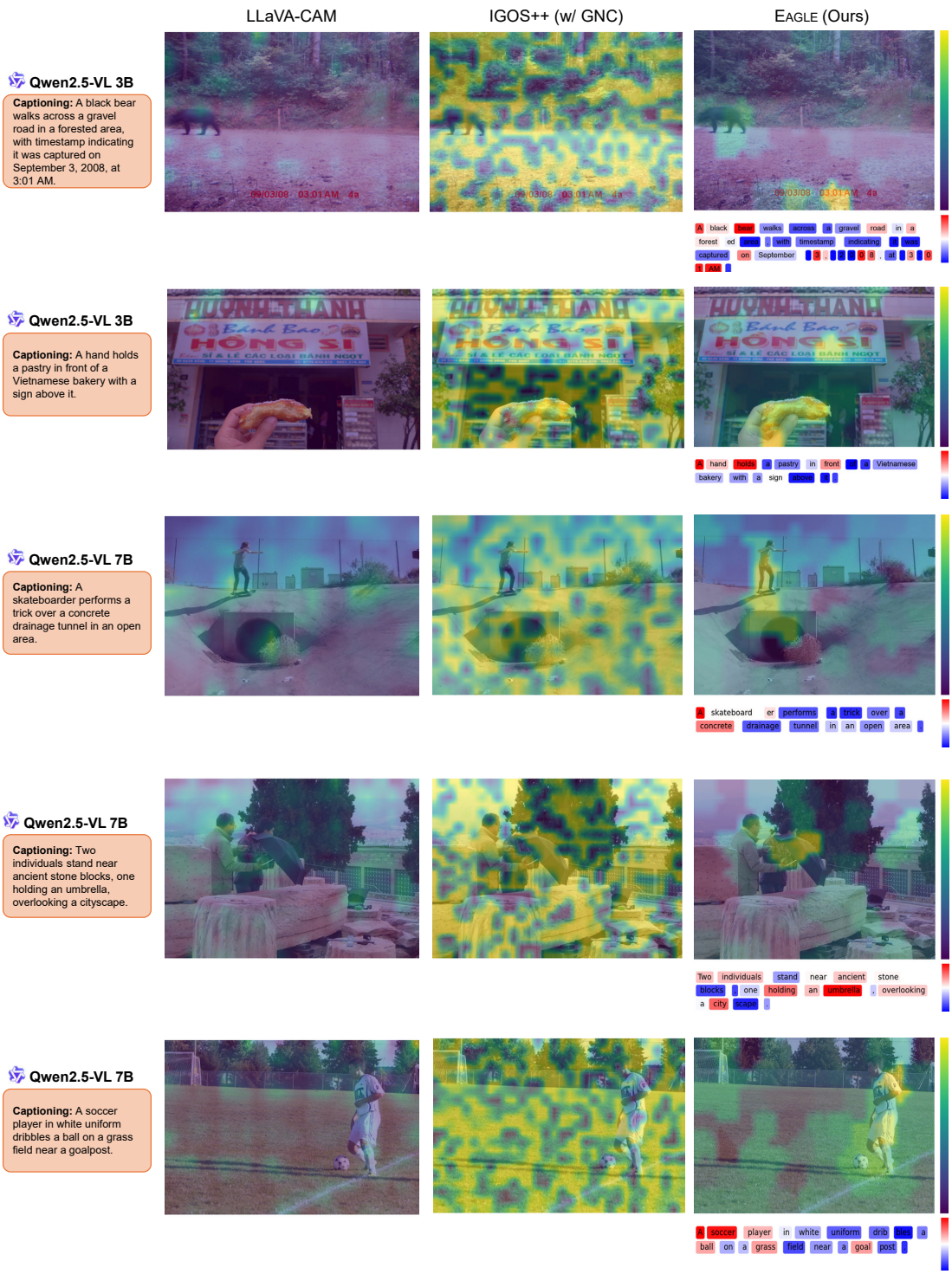

Figure 8. Sentence-level explanation results for **Qwen2.5-VL** on the MS COCO dataset. Our method highlights critical objects with strong correspondence to the generated captions, reducing redundancy in comparison to IGOS++.

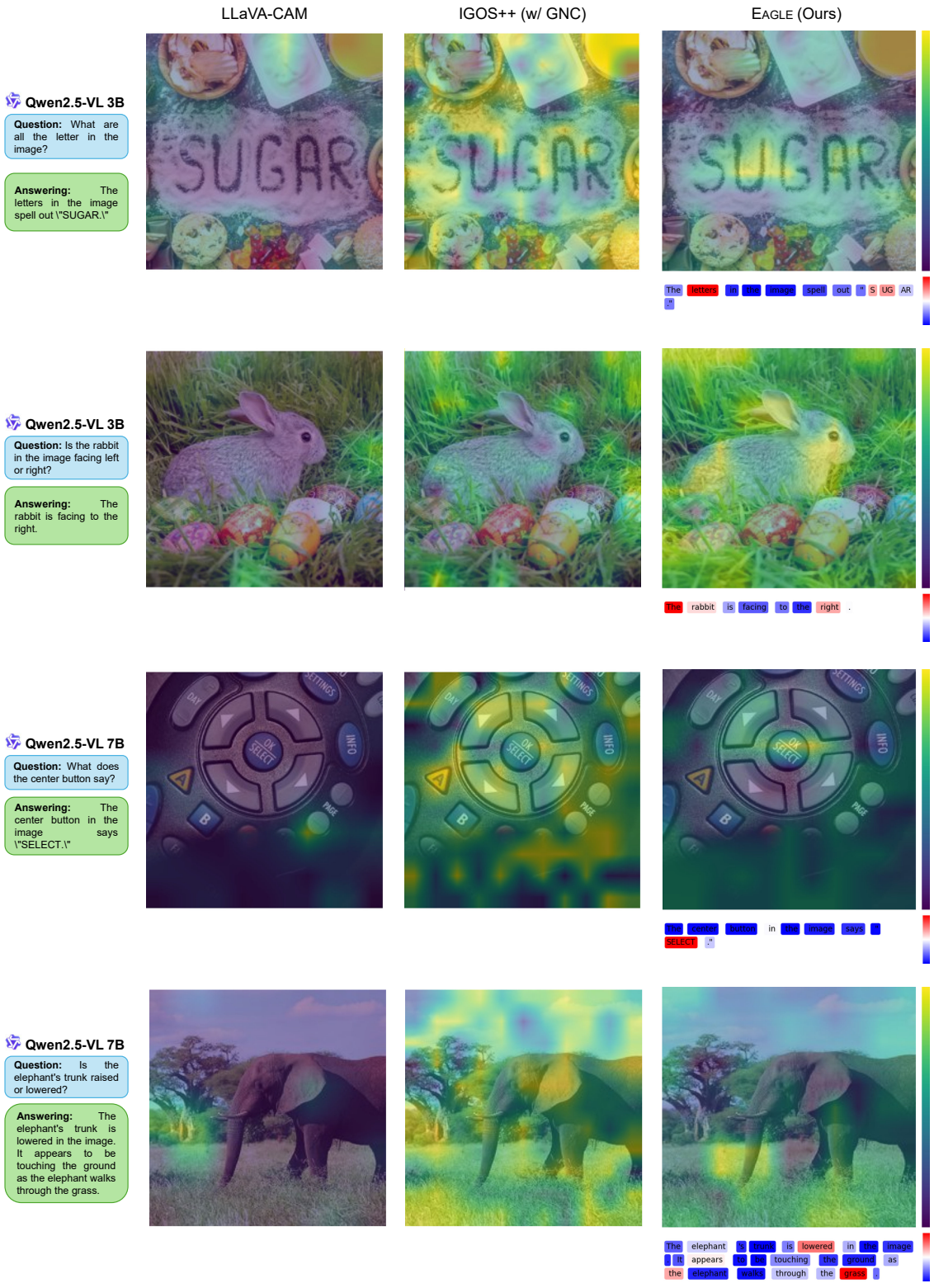

Figure 9. Sentence-level explanation results for **Qwen2.5-VL** on the MMVP dataset. Our method improves alignment between highlighted visual regions and VQA-relevant words, enhancing interpretability.

Figure 10. Sentence-level explanation results for **InternVL3.5** on the MS COCO dataset. Our method captures object-centric regions more consistently than baseline methods.

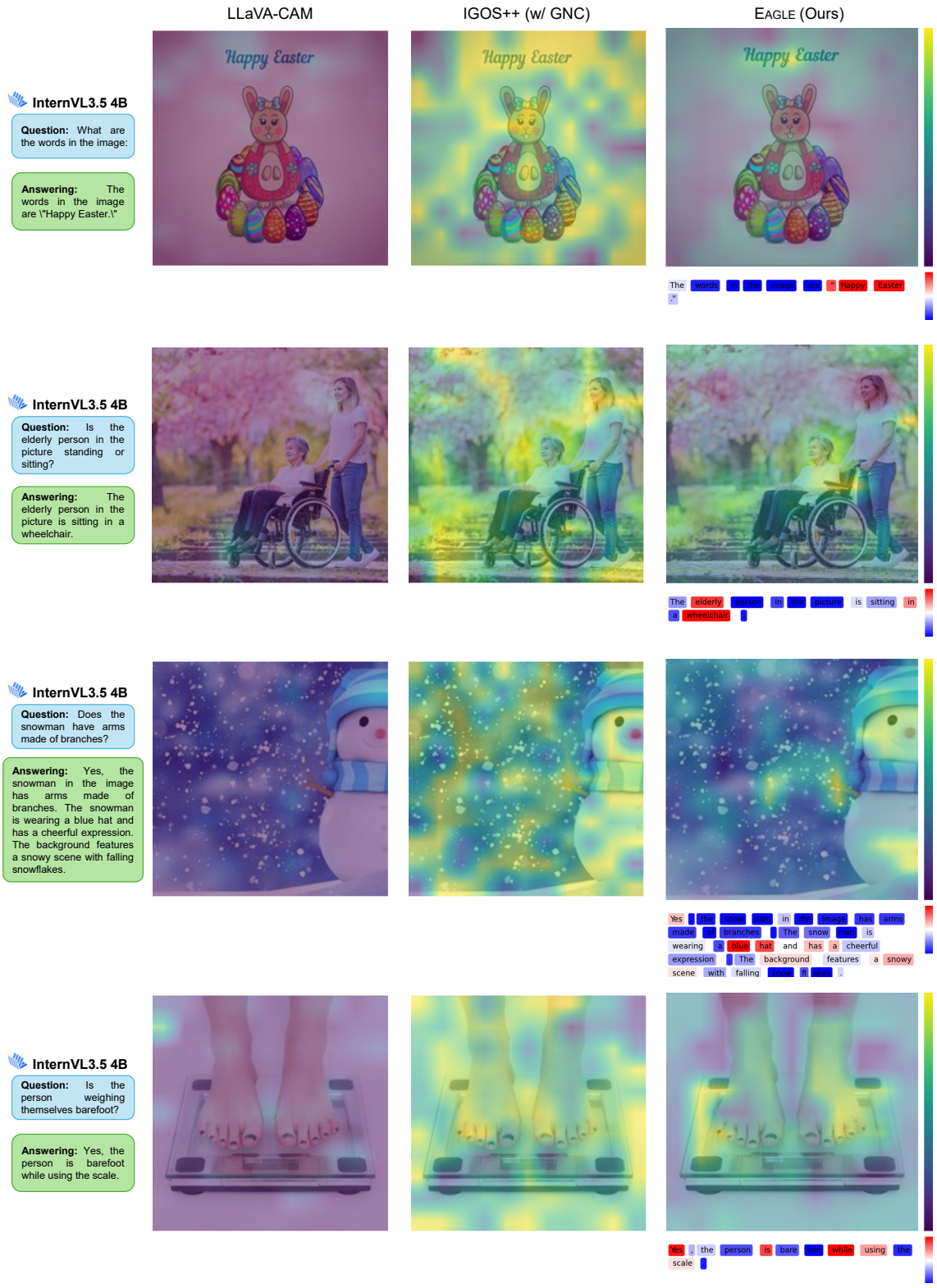

Figure 11. Sentence-level explanation results for **InternVL3.5** on the MMVP dataset. Our approach ensures strong consistency between highlighted evidence and the VQA queries.

Figure 12. Word-level explanation results for **LLaVA-1.5** on the MS COCO dataset. Bounding box overlays show that our method provides sparse yet highly accurate localization.

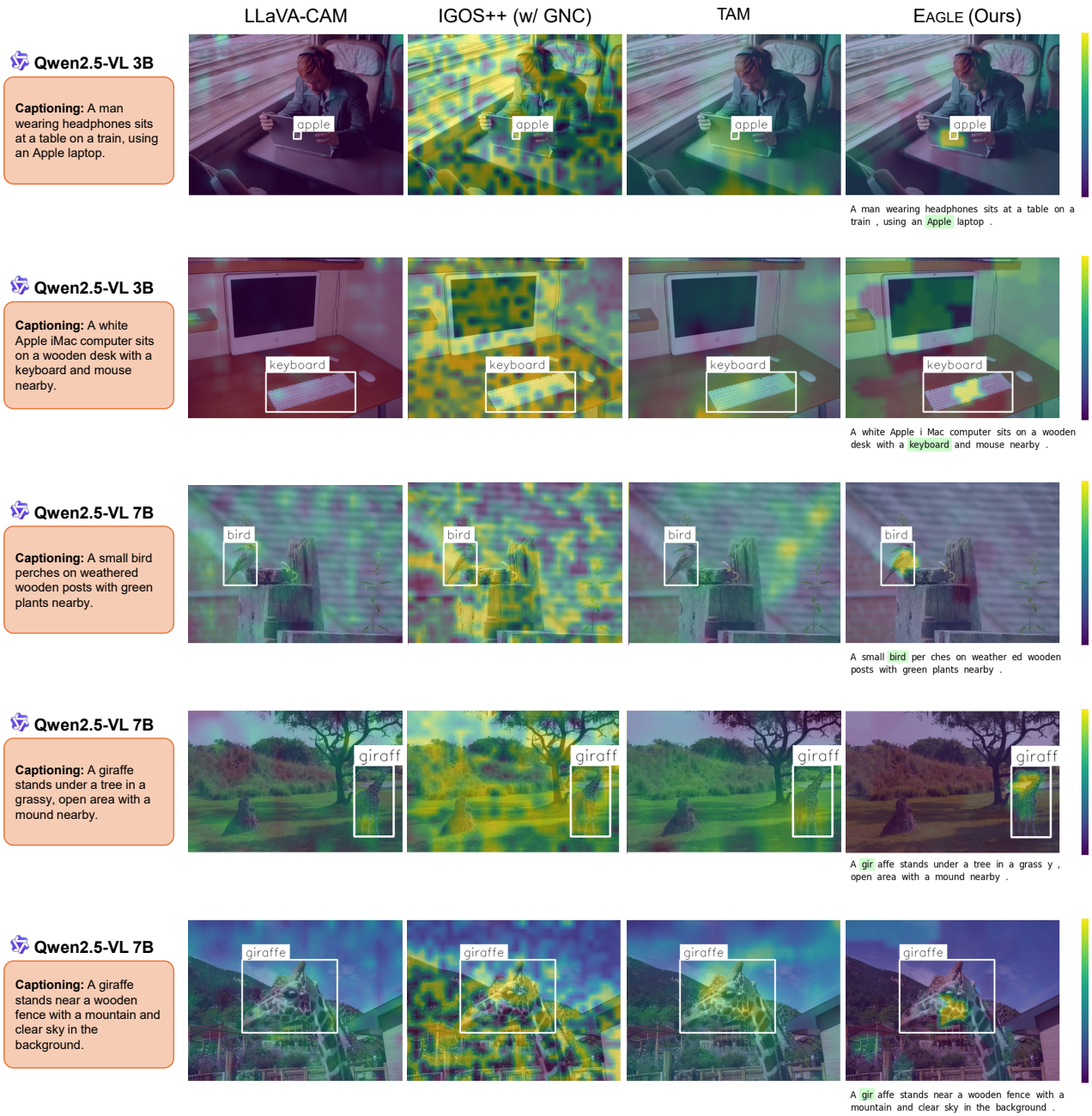

Figure 13. Word-level explanation results for **Qwen2.5-VL** on the MS COCO dataset. Our method produces localized attribution maps with high correspondence to ground-truth bounding boxes.

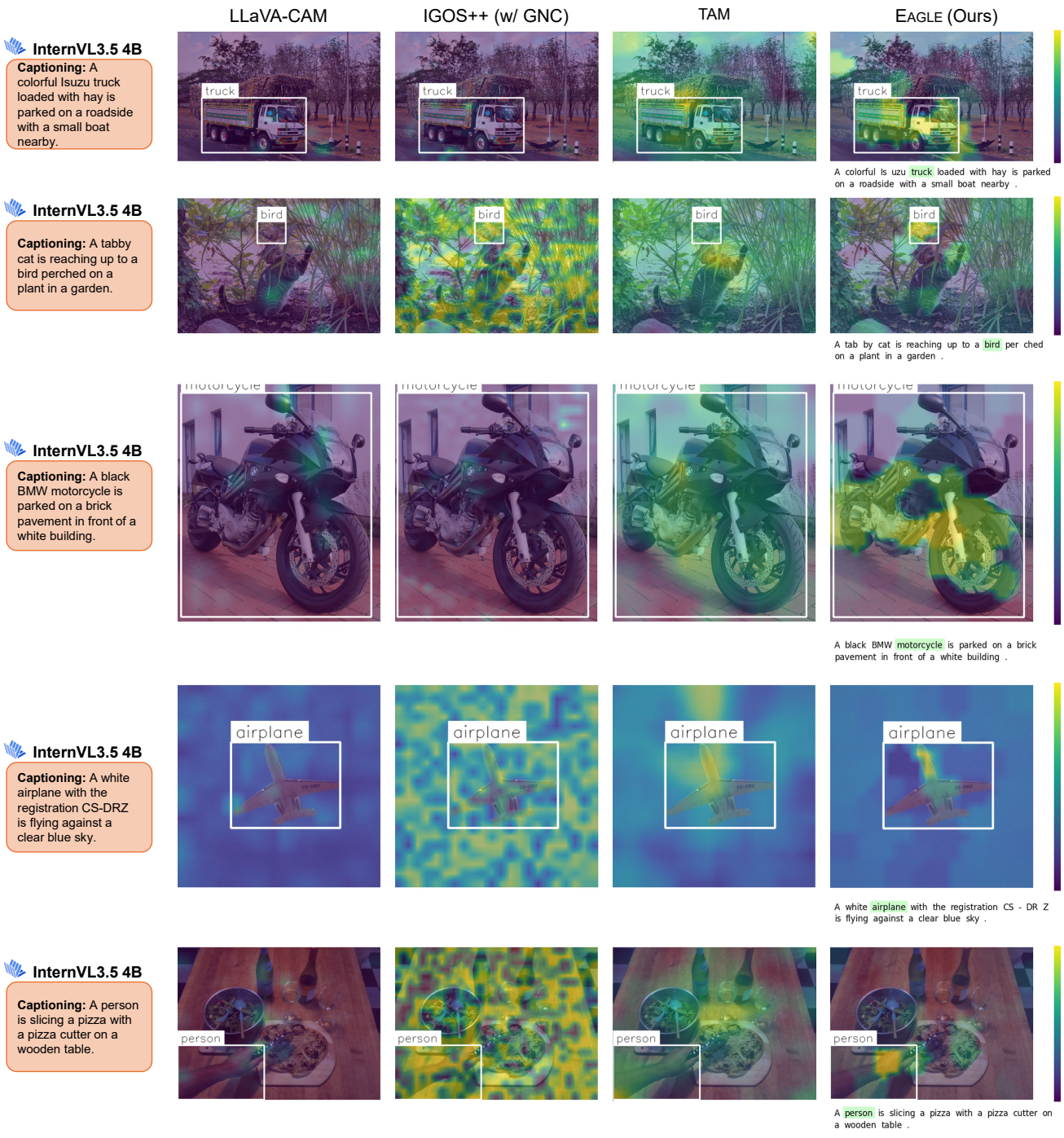

Figure 14. Word-level explanation results for **InternVL3.5** on the MS COCO dataset. Our method captures object-centric highlights with strong correspondence to caption tokens and bounding boxes.

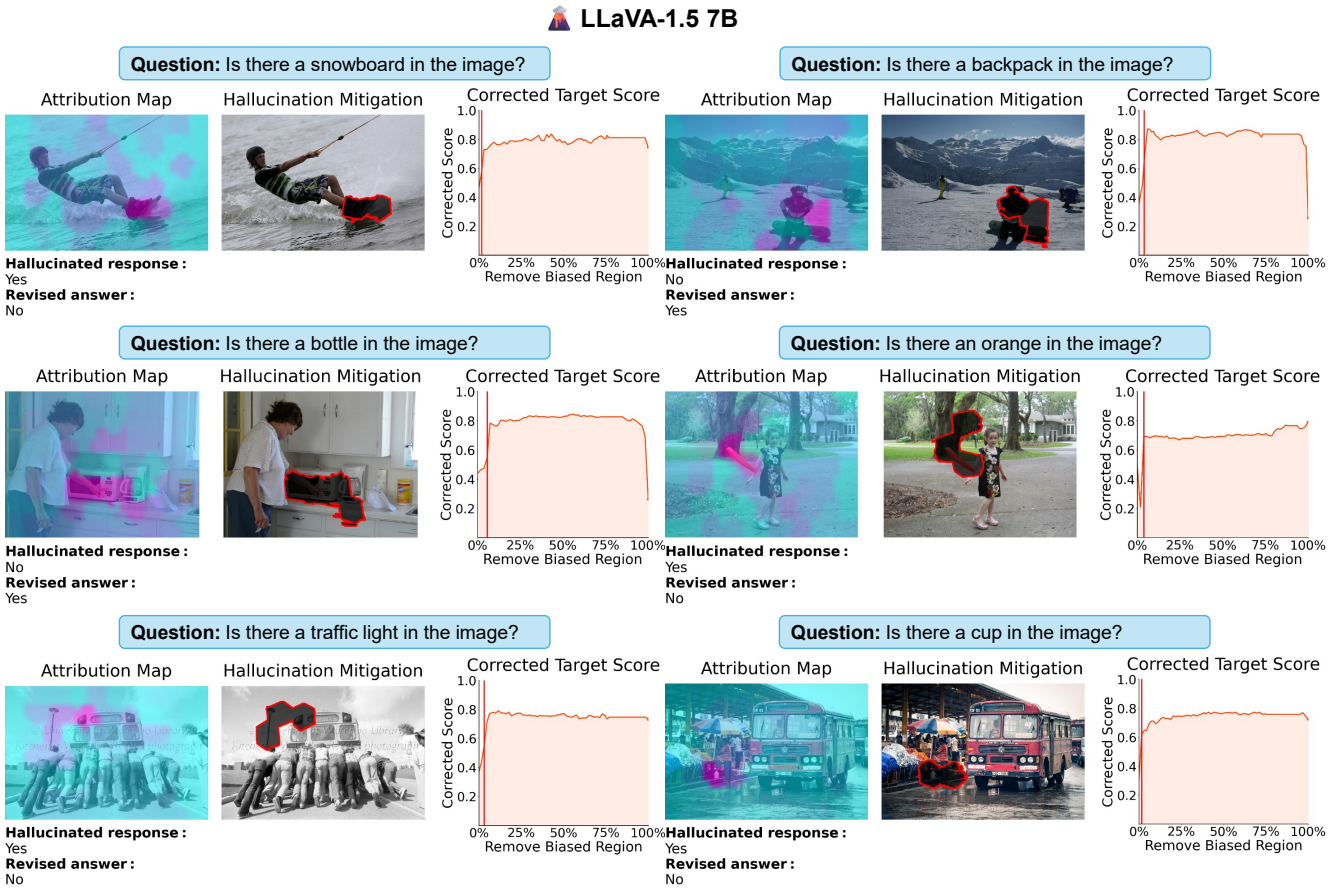

Figure 15. Hallucination attribution for **LLaVA-1.5** on the MS COCO dataset. Our method highlights the minimal hallucination-inducing regions across different queries, such as "snowboard," "traffic light," and "cup."

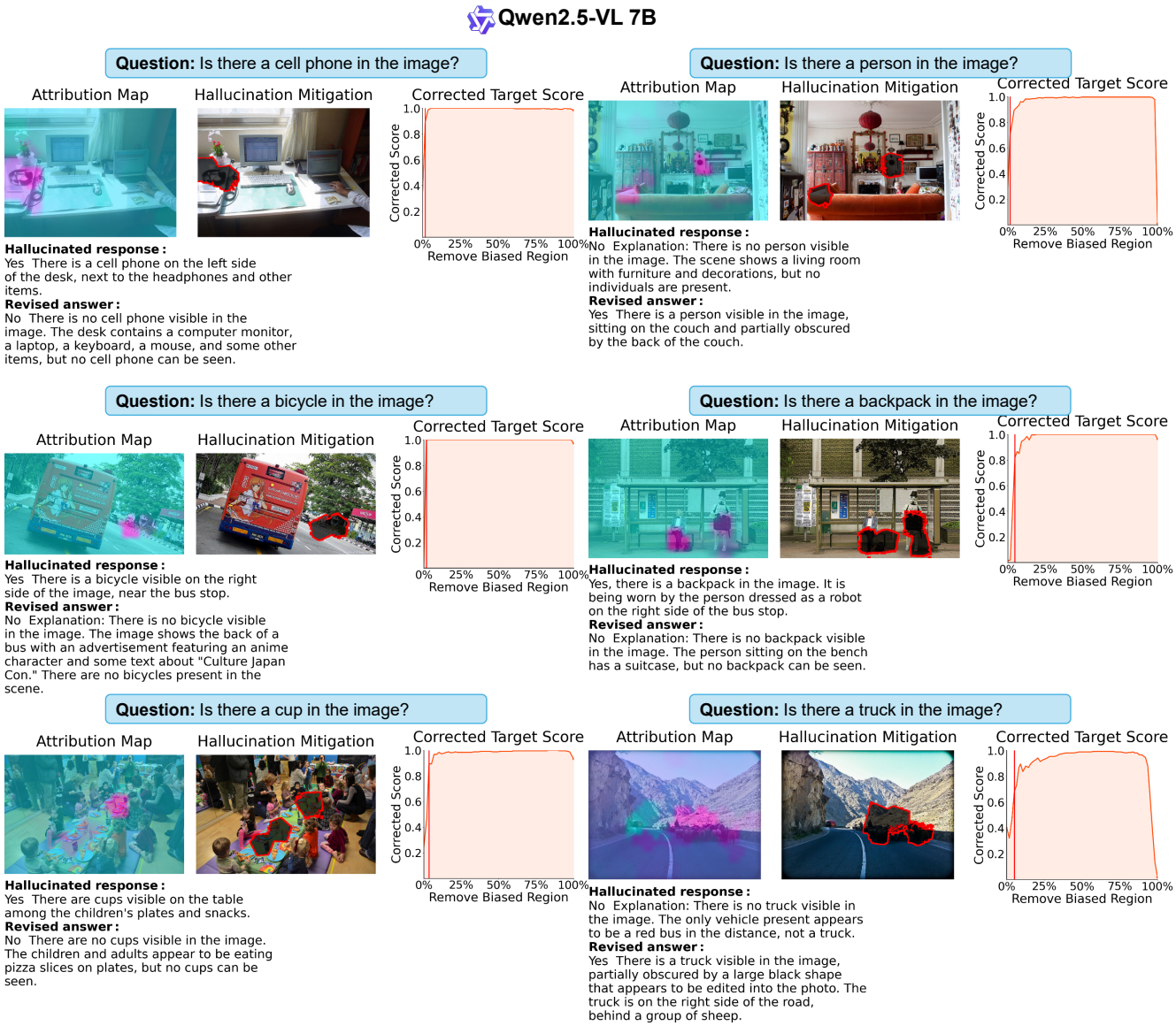

Figure 16. Hallucination attribution for **Qwen2.5-VL** on the MS COCO dataset. Our method isolates misleading cues leading to hallucinations in queries such as "cell phone," "bicycle," and "truck."

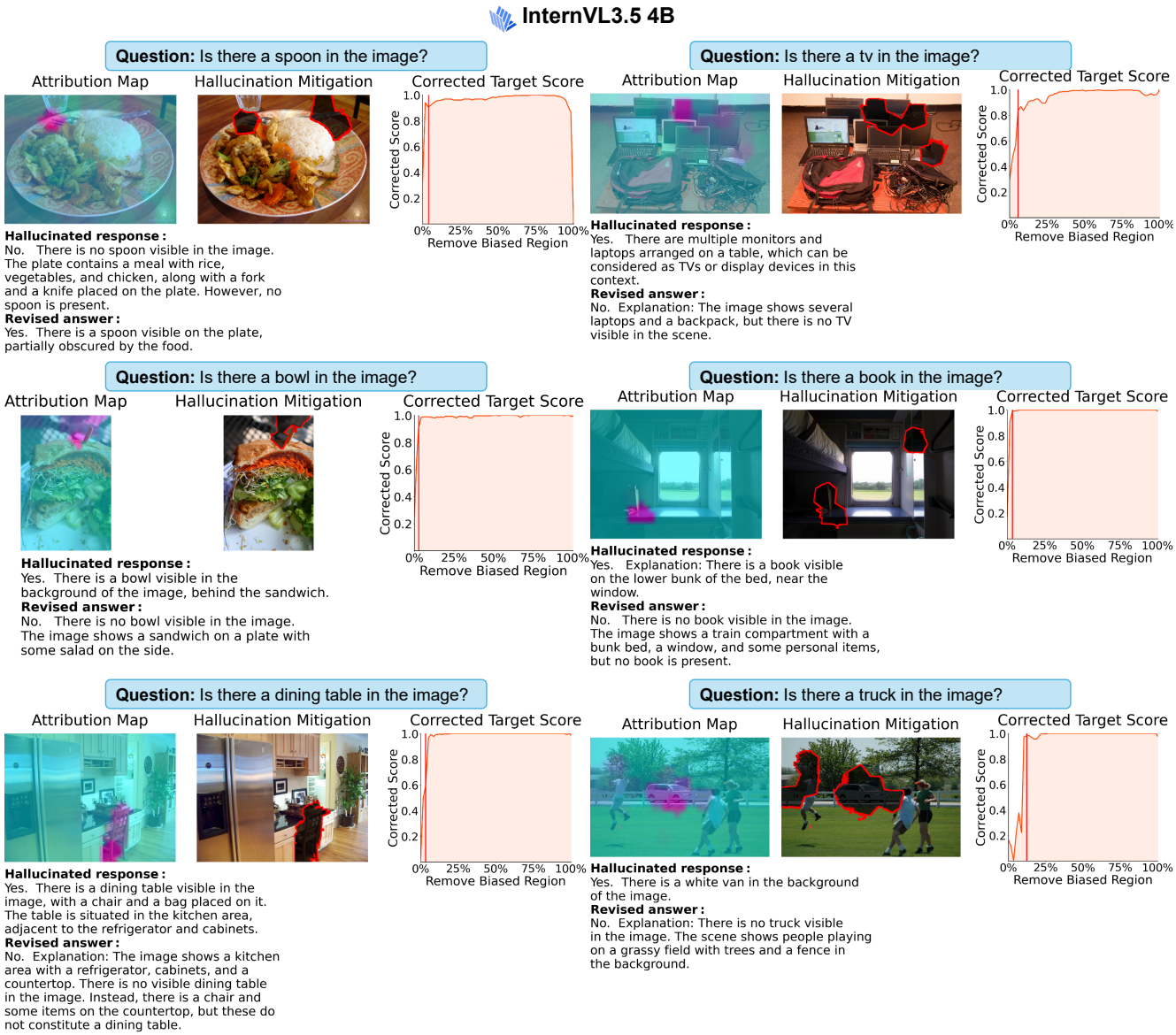

Figure 17. Hallucination attribution for **InternVL3.5** on the MS COCO dataset. Our method identifies hallucination-prone regions for queries such as "spoon," "tv," and "dining table," especially in cases of overlapping or occluded objects.