How Do Medical MLLMs Fail? A Study on Visual Grounding in Medical Images (Supplementary Material for Rebuttal)

Anonymous authors
Paper under double-blind review

A New Figures During Rebuttal

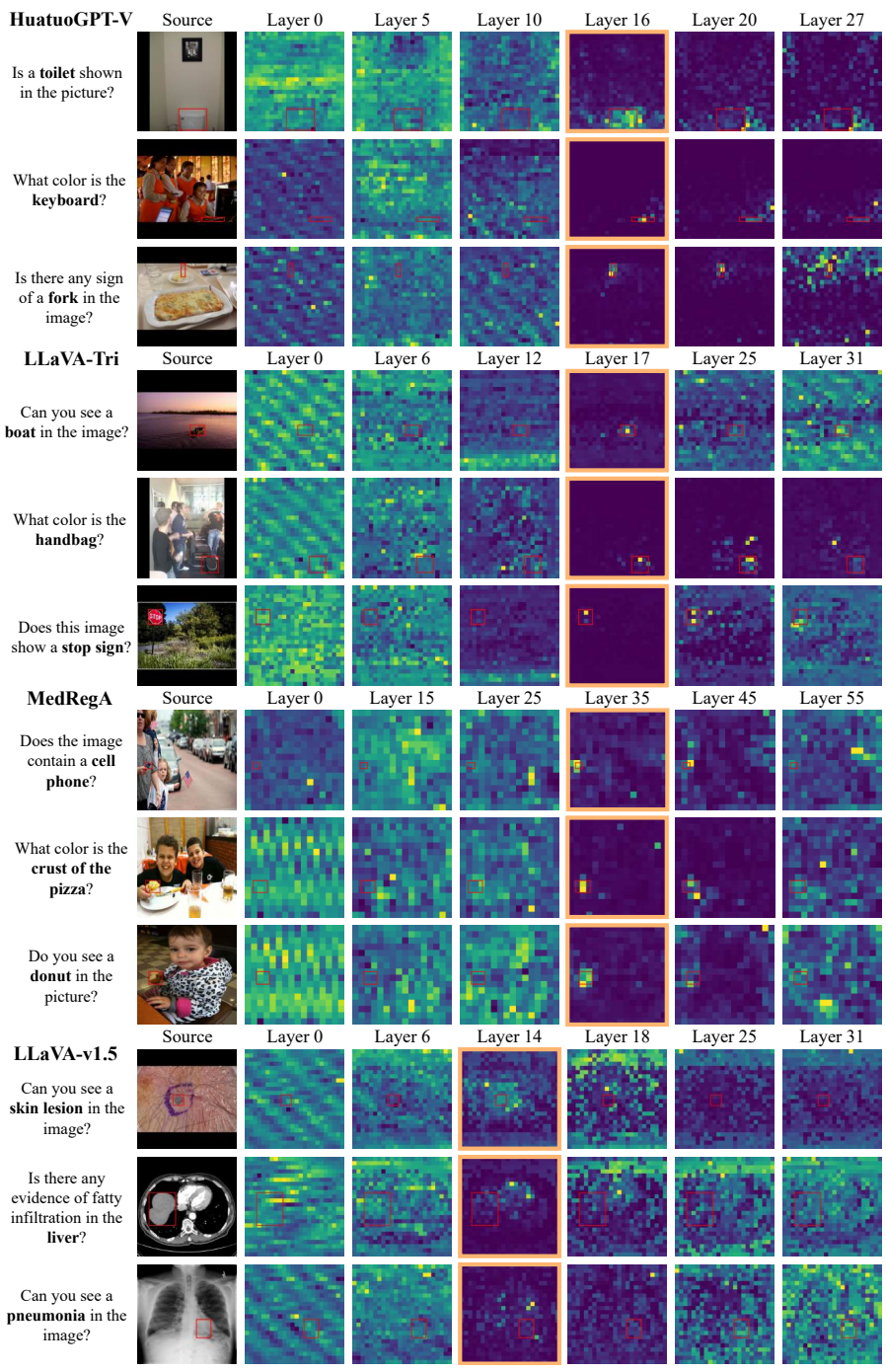

Figure 1: **Qualitative** evaluation of (i) medical MLLMs HuatuoGPT-V, LLaVA-Tri and MedRegA on COCO, and (ii) LLaVA-v1.5 on VGMED. We visualize attention maps across different layers, including those with the lowest KL divergence (highlighted with an orange boundary), which are indicative of layers most relevant to visual grounding in MLLMs. We observe that LLaVA-v1.5 fails to ground predictions in clinically relevant regions when operating on medical images and medical VQA tasks. Furthermore, medical-domain models can ground their predictions when applied to natural images. This is consistent with our quantitative analysis in Fig. 3 of the main paper. Together, they show that medical MLLMs possess good visual grounding capabilities in general-domain settings. **Overall, this confirms that the grounding failure is not due to model weakness, but is fundamentally specific to the medical domain, consistent with our central findings. Inadequate visual grounding is a medical-domain failure mode.**

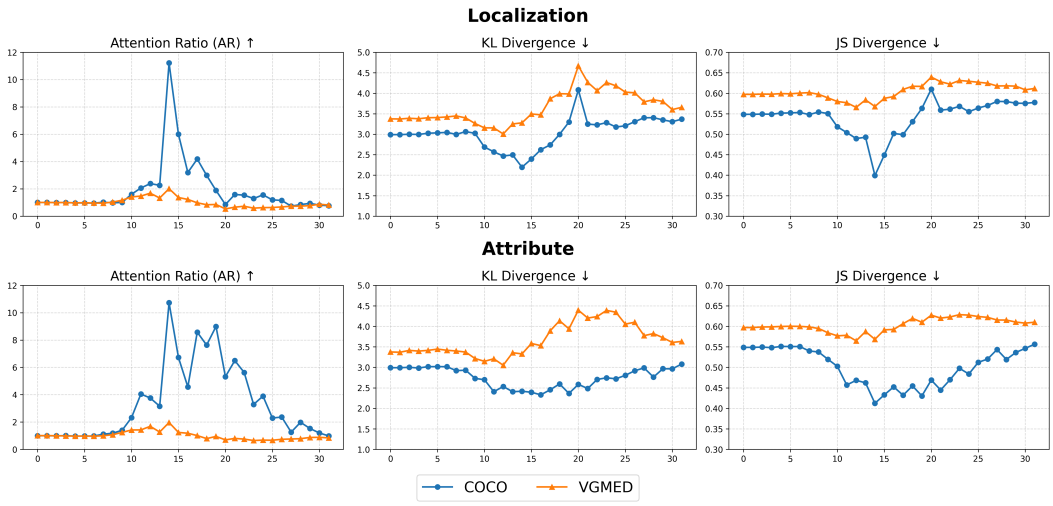

Figure 2: **Quantitative** evaluation of LLaVA-v1.5 on VGMED. We observe that LLaVA-v1.5 fails to ground predictions in clinically relevant regions when operating on medical images and medical VQA tasks.

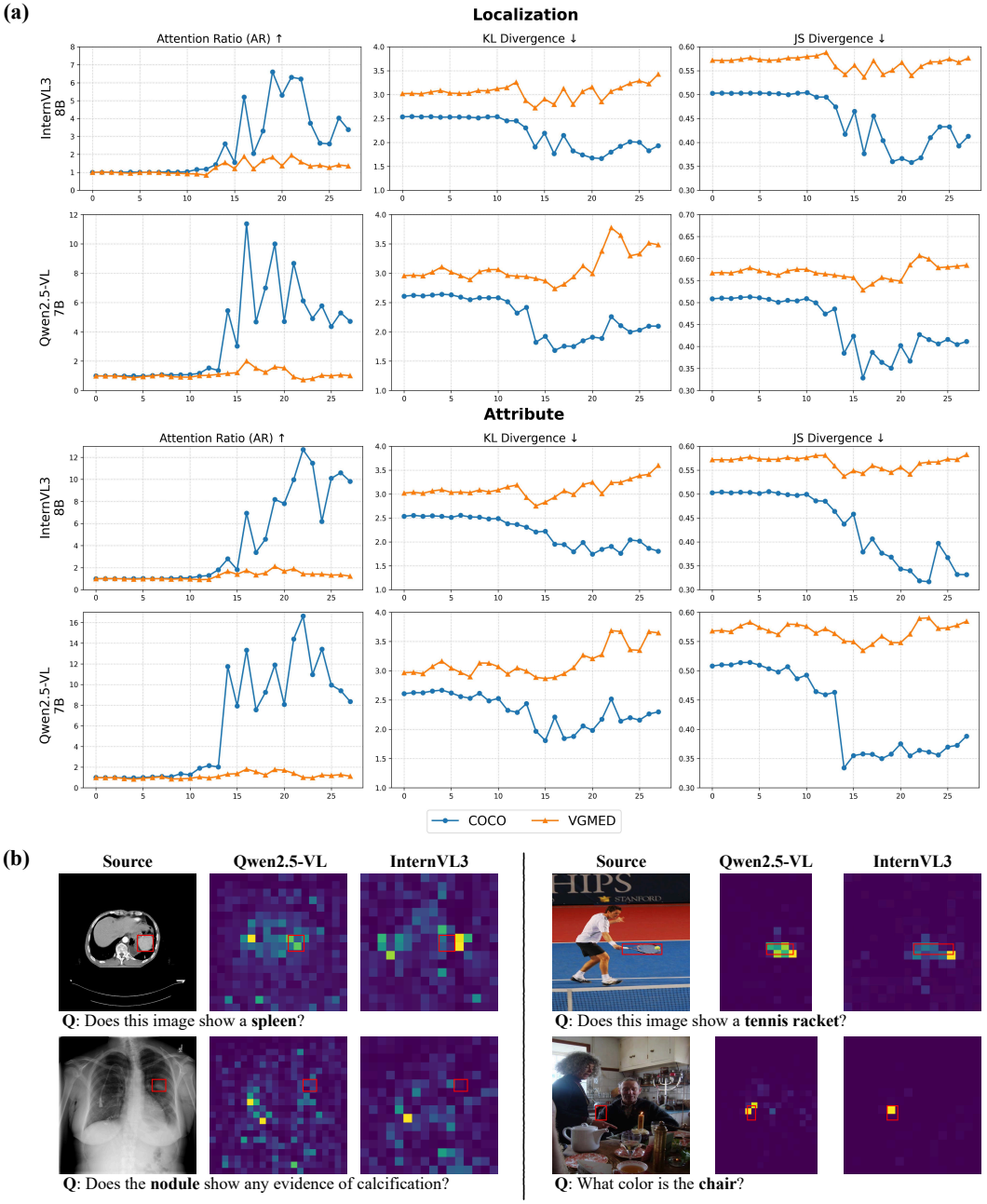

Figure 3: (a) **Quantitative** and (b) **qualitative** evaluation of InternVL3-8B and Qwen2.5-VL-7B on VGMED and COCO. We observe that the visual grounding deficiency in medical domain persists even in these latest general-purpose models.

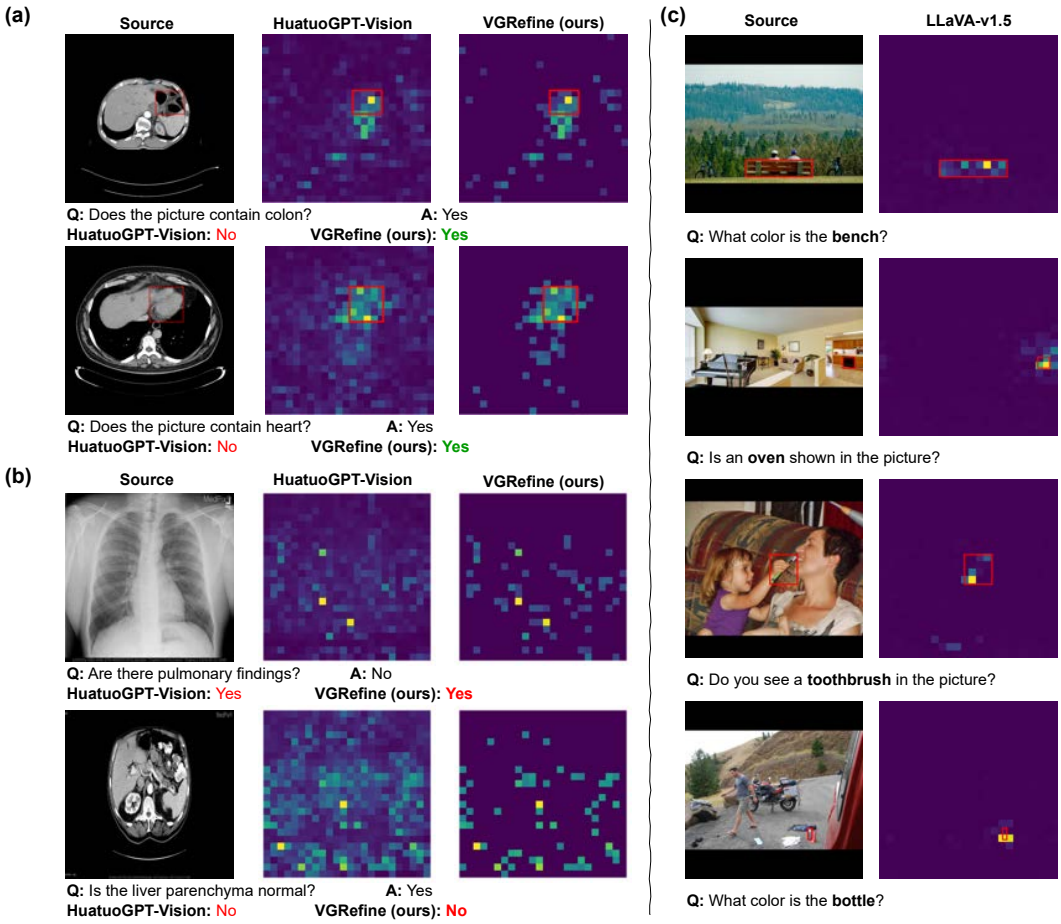

Figure 4: **Representative failure cases of HuatuoGPT-Vision on medical benchmarks.** (a) The model correctly interprets the question but attends to the wrong anatomical region, leading to an incorrect answer. After applying VGRefine, the model's attention shifts toward more clinically relevant region, resulting in the correct prediction. (b) The model misunderstand the question, resulting in both semantic and visual grounding failure. (c) Additionally, we include examples from LLaVA-v1.5 on natural images as a reference of accurate visual grounding. While multiple factors contribute to poor generalization, weak visual grounding consistently emerges as a major and measurable issue, though not the sole cause.

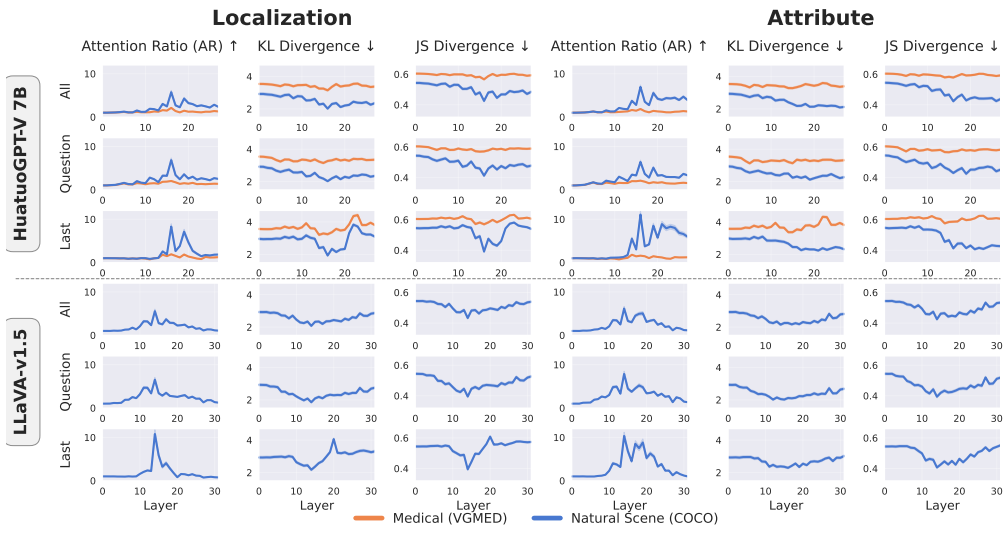

Figure 5: **Comparison of visual grounding when using *all input tokens*, *question-only tokens*, or the *last token* to derive attention maps.** Using two representative MLLMs (HuatuoGPT-V-7B and LLaVA-v1.5), we evaluate how different token-selection strategies affect attention alignment on VGMED and COCO. Across all metrics and layers, attention maps computed from the *last token* achieves equal or better alignment with ground-truth regions compared to the alternative options.

B Clinical Validation During VGMED Curation

As part of the VGMED curation process, clinicians reviewed each sample to verify that (i) the question is properly focused on visual grounding, (ii) it does not require deep or diagnostic-level semantic medical reasoning, and (iii) it remains clinically appropriate and meaningful. An example of the rating interface used during the curation process is shown in Fig. 6.

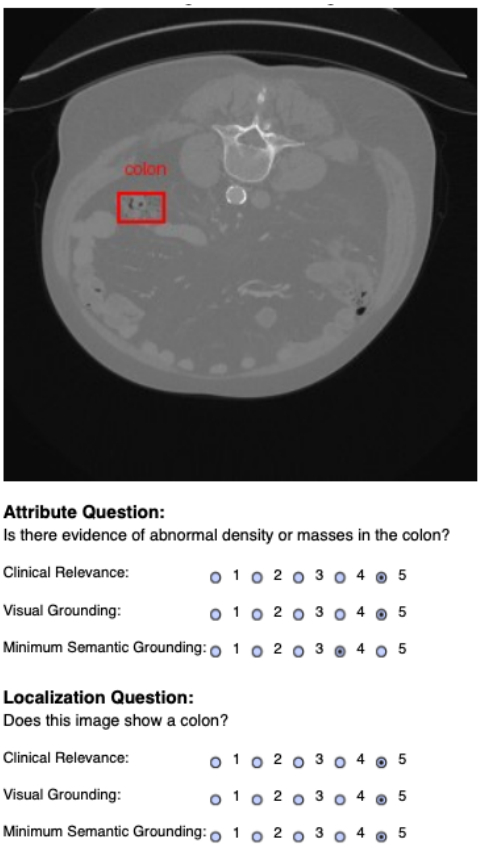

Figure 6: Example of the clinician rating interface used during VGMED curation.

Clinical Relevance

- **1:** Irrelevant or misleading; the question is clinically inappropriate or nonsensical in this context.
- **2:** Marginally relevant; the question has limited medical value or loosely pertains to the case.
- **3:** Acceptable; the question is reasonable in clinical significance.
- **4:** Clinically useful; the question is clearly relevant and meaningful to medical interpretation.
- **5:** Highly relevant and valid; the question is well-phrased, accurate, and directly supports clinical reasoning.

Visual Grounding

- **1:** It refers to other anatomy or ignores the boxed area entirely; ignores the region.
- **2:** The question has only a weak or incidental connection to the boxed region; the area is largely irrelevant to the text.
- **3:** It reasonably overlaps or implies the boxed region.
- **4:** Clear reference to the boxed region.
- **5:** Perfectly aligned, the question precisely refers to the boxed region.

Minimum Semantic Grounding

- **1:** Very deep semantic grounding; requires advanced, multi-step clinical reasoning, such as staging, prognosis, mechanisms, or treatment decisions.
  Examples:
  "What is the appropriate treatment for this condition?"
  "How does this imaging pattern affect the patient's prognosis?"
- **2:** High semantic grounding; requires reasoning about specific diseases or well-defined diagnostic entities. Substantial medical knowledge is needed.
  Example:
  "What diseases are included in the image?"
- **3:** Moderate semantic grounding; requires linking features to broad categories of pathology, such as distinguishing between growth, inflammation, or degeneration.
  Example:
  "Do the changes suggest a long-standing damage?"
- **4:** Low–moderate semantic grounding; requires recognition of more specific medical descriptors, but does not involve broad pathology categories or diagnostic reasoning.
  Examples:
  "Does the structure appear to be pushing against or displacing nearby tissues?"
  "Is there a region that appears more diffuse rather than well-demarcated?"
- **5:** Low semantic grounding requires only basic clinical or anatomical recognition (e.g., body parts, organs, simple structures, fractures, nodules).
  Examples:
  "Does the bone show a visible fracture line?"
  "Is there a nodule in this region?"

Therefore, a rating of 3 represents acceptable threshold across all three dimensions: the sample is clinically relevant, visually grounded, and does not require deep semantic knowledge.

During the benchmark curation process, all samples receiving any score below 3 were discarded. Consequently, every VGMED sample satisfies 3 or above on all criteria. This ensured that retained samples genuinely test visual grounding rather than medical reasoning.

Furthermore, as summarized in Tab. 1, the vast majority of clinician ratings are in the upper categories (4–5), with only a minor proportion of samples receiving a rating of 3 across any axis.

Table 1: Percentage distribution of clinician ratings (3–5) across all axes for Attribute and Localization questions.

Type	Category	Rating 3 (%)	Rating 4 (%)	Rating 5 (%)
Attribute	Clinical Relevance	3.31	4.11	92.58
	Min. Semantic Grounding	0.37	10.38	89.25
	Visual Grounding	4.04	12.18	83.77
Localization	Clinical Relevance	0.02	0.52	99.46
	Min. Semantic Grounding	0.05	5.76	94.19
	Visual Grounding	3.96	11.79	84.25

REFERENCES