How Do Medical MLLMs Fail? A Study on Visual Grounding in Medical Images

Anonymous authors
Paper under double-blind review

Abstract

Generalist multimodal large language models (MLLMs) have achieved impressive performance across a wide range of vision-language tasks. However, their performance on medical tasks—particularly in zero-shot settings where generalization is critical—remains suboptimal. A key research gap is the limited understanding of why medical MLLMs underperform in medical image interpretation. **In this work**, we present a pioneering systematic investigation into the visual grounding capabilities of state-of-the-art medical MLLMs. To disentangle *visual grounding* from *semantic grounding*, we design VGMED, a novel evaluation dataset developed with expert clinical guidance, explicitly assessing the visual grounding capability of medical MLLMs. We introduce new quantitative metrics and conduct detailed qualitative analyses. Our study across **eight** state-of-the-art (SOTA) medical MLLMs validates that they often fail to ground their predictions in clinically relevant image regions. We note that this finding is specific to medical image analysis; in contrast, prior work has shown that MLLMs are capable of grounding their predictions in the correct image regions when applied to natural scene images. Motivated by these findings, we propose VGRefine, a simple yet effective inference-time method that refines attention distribution to improve visual grounding in medical settings. Our approach achieves SOTA performance across 6 diverse Med-VQA benchmarks (over 110K VQA samples from 8 imaging modalities) without requiring additional training or external expert models. Overall, our work, for the first time, systematically validates inadequate visual grounding as one of the key contributing factors for medical MLLMs' under-performance. Code and additional experiments are included in the Supp.

1 Introduction

Generalist multimodal large language models (MLLMs) have demonstrated strong performance across a broad range of vision-language tasks, including visual question answering (VQA) (Wang et al., 2024; Dai et al., 2023; Liu et al., 2023; Chen et al., 2024b; Liu et al., 2024b), image captioning (Li et al., 2023b; Wu et al., 2024), science and mathematical reasoning (Liu et al., 2024d; Zhuang et al., 2025; Shi et al., 2024). Recent efforts have extended these models to the medical domain, with the goal of developing medical MLLMs that can leverage their generalization capabilities to support diverse clinical decision-making tasks.

Medical MLLMs. Recent work has explored extending general-purpose MLLMs to the medical domain, with many approaches focusing on constructing multimodal medical datasets and incorporating external expert models. In Li et al. (2023a), a large-scale biomedical figure-caption dataset is built from PubMed Central to fine-tune LLaVA, resulting in LLaVA-Med. However, its performance in zero-shot settings remains suboptimal and heavily reliant on dataset-specific fine-tuning. HuatuoGPT-Vision (Chen et al., 2024a) leverages GPT-4V to construct a large image-text dataset with refined annotations, but also lacks strong zero-shot generalization. VILA-M3 (Nath et al., 2024) incorporates external medical expert models to assist medical image analysis tasks. Recently, in Xie et al. (2025), the authors introduce MedTrinity-25M, a dataset comprising 25 million medical images, and propose LLaVA-Tri, a model pretrained on this dataset to improve regional focus in medical images. (See Supp. for additional review of related work.)

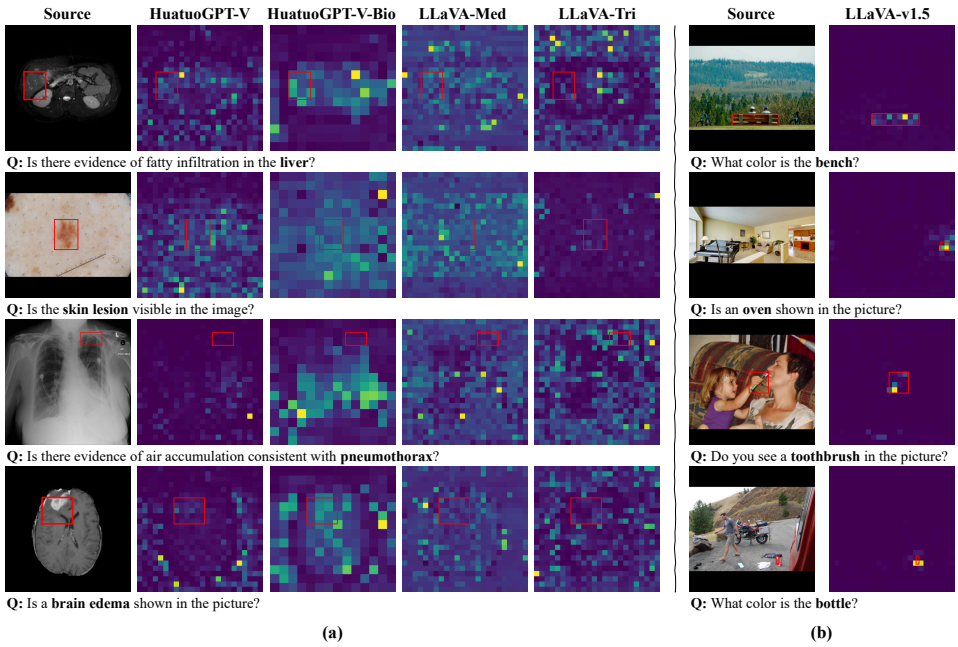

Figure 1: **Visual grounding issues in state-of-the-art medical MLLMs.** (a) Column 1 shows input medical images with expert-annotated ground-truth regions (red boxes). Columns 2–5 display attention distributions from representative medical MLLMs. (b) Column 1 shows natural scene images with annotated ground-truth bounding boxes, and column 2 shows attention distributions from LLaVA-v1.5. For the first time, we systematically validate that state-of-the-art medical MLLMs often suffer from *inadequate visual grounding*—they fail to accurately localize and interpret image regions that are clinically relevant to the question. We note that, in contrast, when applied to natural images, MLLMs are capable of grounding their predictions in the correct image regions (Zhang et al., 2025a). Attention maps are taken from the LLM layers identified as most relevant to visual grounding (see Sec. 2 for details).

Despite these advances, most existing medical MLLMs strongly rely on training or fine-tuning with samples from downstream datasets. They continue to underperform on medical VQA tasks in the zero-shot setting—where no downstream task samples are seen during training or fine-tuning—thus falling short of the goal of developing truly generalist medical MLLMs. This raises a key question: *Why do medical MLLMs struggle with medical image interpretation, despite their success in general-domain tasks?*

Research Gap. There remains a lack of deeper analysis into the underlying causes of medical MLLMs' suboptimal performance in the important zero-shot setting. Particularly, there is a lack of studies to systematically examine the internal failure modes of these models—particularly in terms of *how* and *where* predictions are derived from visual inputs. Without such analysis, it remains unclear whether performance limitations stem from a lack of clinical task understanding (semantic grounding) or from an inability to accurately localize and interpret relevant image regions (visual grounding). Advancing our understanding of these failure modes is essential for building robust generalist medical MLLMs for real-world clinical deployment.

Our work underscores the importance of explicitly distinguishing between *semantic grounding* (Lu et al., 2024; Lyre, 2024) and *visual grounding* (Xiao et al., 2024) in medical tasks. This distinction is particularly critical for Med-VQA, which—unlike general-domain VQA—often requires deep domain-specific reasoning. For example, answering a question like "What diseases are included in the image?" requires the model to reason about the anatomical structures and visual features that are relevant to specific pathologies. A model may experience *failure in semantic grounding*—that is, it lacks the medical knowledge to determine *what* to look for. Alternatively, it may experience *failure in visual grounding*—it cannot accurately *localize and interpret* the relevant regions in the medical image, even when it knows what to look for. As medical MLLMs increasingly incorporate large-scale biomedical knowledge to enhance semantic grounding, we argue that visual grounding may emerge as the primary bottleneck limiting further progress.

Our Contribution. In this work, we present a pioneering systematic investigation aimed at advancing the understanding of failure modes and the visual grounding capabilities of medical MLLMs (Fig. 1). To disentangle visual grounding from semantic grounding, we co-create a novel evaluation dataset with 3 clinicians, named VGMED, a dataset for Visual Grounding analysis of MEDical MLLMs. VGMED ensures focused evaluation of whether MLLMs can accurately localize and interpret the relevant regions in medical images. We introduce new quantitative metrics and qualitative analyses to assess visual grounding performance—that is, the extent to which model predictions are grounded in clinically relevant visual evidence. *Critically, by using VGMED to evaluate **eight** SOTA medical MLLMs, we reveal for the first time that even the most advanced models frequently rely on spurious or irrelevant regions, highlighting inadequate visual grounding as a pervasive and fundamental failure mode.* We note that this finding is specific to medical image analysis; in contrast, prior work has shown that MLLMs are capable of grounding their predictions in the correct image regions when applied to natural images (Zhang et al., 2025a).

To address this, we propose VGRefine, a simple yet effective inference-time method that improves visual grounding by refining internal attention distributions. VGRefine requires no additional training. Across 6 diverse Med-VQA benchmarks, comprising over 110K VQA samples from 8 imaging modalities (CT, MRI, X-ray, OCT, dermoscopy, microscopy, fundus, ultrasound), VGRefine consistently achieves improved and SOTA performance. Overall, our work offers new insights into the failure modes of medical MLLMs and establishes visual grounding analysis as a necessary diagnostic tool for advancing medical MLLMs in clinical applications.

2 Investigation of Visual Grounding in Medical MLLMs

Despite recent advances, medical MLLMs continue to underperform on complex medical image reasoning tasks, particularly in medical VQA (Hu et al., 2024; Jeong et al., 2024). In this work, we conduct a systematic study to validate that a key limitation lies in inaccurate visual grounding. As a starting point, we analyze attention maps from the model layers most relevant to visual grounding (details on layer selection are provided in Sec. 2.4). As shown in Fig. 1, for medical images, MLLMs' attentions often fail to align with clinically relevant regions.

2.1 A New Dataset for Visual Grounding Analysis

Existing medical VQA datasets are ill-suited for visual grounding analysis. To rigorously evaluate medical MLLMs' visual grounding, we aim to systematically assess the extent to which their outputs are supported by clinically relevant regions of the image (e.g., organs, tissues, or lesions essential for answering a given question). However, existing medical VQA datasets are ill-suited for this purpose, as illustrated in Fig. 2 (a). Many questions, such as "*What diseases are included in the picture?*", can be answered without referencing specific image regions. In contrast, questions like "*What diseases are included in the picture?*" require substantial medical knowledge to determine what to look for, since different diseases, including their stages or subtypes, can manifest with varied and often subtle visual patterns. These patterns are not always well-documented in text and may depend on clinical interpretation, making it difficult to determine whether model failures stem from inadequate semantic grounding or from visual grounding alone.

VGMED: A new dataset for Visual Grounding analysis of MEDical MLLMs co-created with clinicians. To address this gap, we build an evaluation dataset VGMED, focusing on visual grounding analysis, as illustrated in Fig. 2 (b). VGMED was co-created with three certified medical doctors (general practice, neurology, radiology) to ensure annotation accuracy and clinical relevance, including two senior clinicians with over ten years of experience. One expert also serves as Director (AI and Data Science) at a national medical center. Their contributions included: (1) co-designing GPT prompts to elicit clinically meaningful and visually grounded questions, (2) reviewing and refining all samples for clinical relevance and grounding focus, and (3) verifying that all samples require reference to the annotated region.

Our VGMED dataset is constructed from over 40 publicly available medical image segmentation datasets, with detailed information summarized in Table C.2. The original segmentation masks are converted into bounding boxes to support visual grounding analysis. To ensure diversity across imaging modalities and anatomical regions, we filter 13,962 samples, each consisting of a medical

image paired with a ground-truth bounding box. The distributions of modalities and body parts are illustrated in Fig. 2 (b).

For each image–bbox pair, we construct clinically meaningful questions that target specific anatomical or pathological regions, guided by input from clinical experts. This allows us to conduct fine-grained visual grounding analysis. The questions are first generated using GPT-4 and subsequently reviewed and validated by medical professionals. They fall into two categories: *localization* and *attribute* questions. Localization questions inquire about the presence or identification of a specific organ or lesion, whereas attribute questions focus on visual properties such as size, shape, or abnormality (see Fig. 2 for details). GPT-4 is prompted to ensure that questions are both clinically relevant and visually grounded, requiring attention to the entire annotated region. In total, our dataset contains approximately 28K image–bbox–question triplets.

As the reference point, we randomly draw the same number of samples from MS COCO (Lin et al., 2014), using the same question generation pipeline for the evaluation of natural scene images. We include all prompts used in localization and attribute questions generation in Supp I.

Remark: Co-created with 3 clinicians, VGMED is a dataset for *evaluation and analysis* of visual grounding in medical domain. The size of VGMED (28K samples) is comparable to datasets typically used in general-domain visual grounding evaluation and studies (see Supp C).

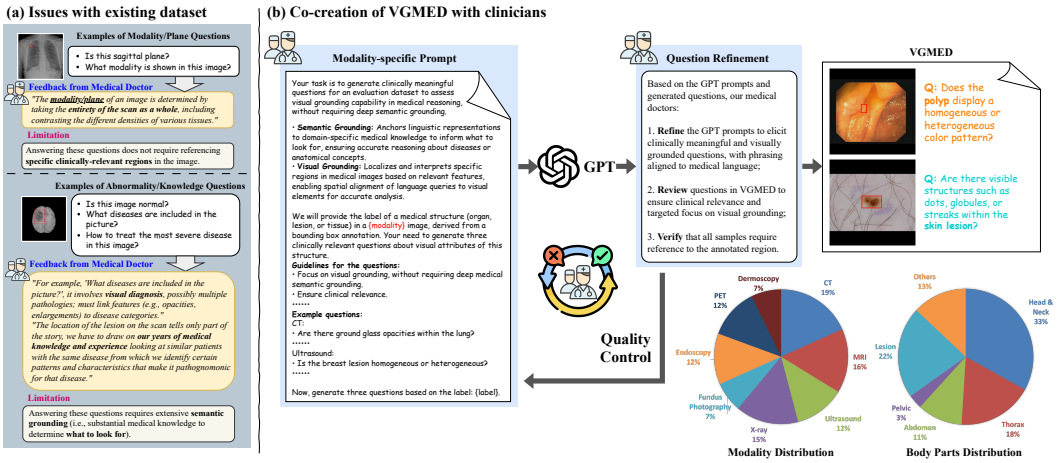

Figure 2: **Co-creation of VGMED with clinicians for visual grounding assessment.** Existing Med-VQA datasets often include questions about image modality or plane, which can be answered without referencing specific image regions. They also contain many abnormality- or knowledge-based questions that require substantial medical expertise to determine what to look for. As a result, existing datasets are not well-suited for analyzing visual grounding. In contrast, our dataset leverages LLM prompting and clinical expert guidance to generate clinically meaningful localization and attribute questions that are explicitly grounded in annotated image regions, enabling rigorous assessment of the visual grounding capabilities of medical MLLMs. **Best viewed in color and with zoom.**

2.2 Quantifying MLLMs' Visual Grounding with Attention Maps

Measuring MLLMs' visual grounding. To evaluate how multimodal large language models (MLLMs) ground their predictions in visual evidence, we analyze internal attention maps that indicate which image regions the model attends to. Attention maps are widely used in recent studies to evaluate visual grounding in general-domain MLLMs (Zhang et al., 2025a; Kang et al., 2025; Kaduri et al., 2024). Importantly, Zhang et al. (2025a) demonstrated that attention distributions can reliably capture visual grounding in natural scene images. This enables us to directly compare the visual grounding in medical images and natural scene images.

Alternative grounding indicators, such as gradient-based saliency and causal perturbation, are in principle applicable but are considerably more expensive at scale. Gradient-based saliency methods (Selvaraju et al., 2017; Ismail et al., 2021) require backpropagation for each input, making them substantially more computation-intensive than directly using attention maps from the forward pass.

Causal perturbation techniques (Fong & Vedaldi, 2017; Hooker et al., 2019) demand a new forward pass for each perturbed input (e.g., region masking or token removal), which may quickly become prohibitive for large-scale medical grounding analysis.

Attention maps in MLLMs. We extract cross-attention weights from the last input text token to each of the N^2 image tokens across all L layers and H attention heads of the LLM (Zhang et al., 2024; Kang et al., 2025). For each layer ℓ and head h, we denote the attention vector as $\alpha^{\ell,h} \in \mathbb{R}^{N^2}$, and compute the average across heads to obtain a per-layer attention map $A^\ell = \frac{1}{H}\sum_{h=1}^{H} \alpha^{\ell,h}$. Then we reshape A^ℓ into a spatial attention map of size $N \times N$.

Attention Ratio (AR). We aim to measure the alignment from the model's attention map to the ground truth bounding box. For this purpose, we apply attention ratio (AR), defined as the sum of attention inside the ground truth bounding box divided by the average attention inside the bounding box of the same size (Zhang et al., 2025a). Let $A \in \mathbb{R}^{N\times N}$ denote the attention map over image patch tokens, and let $M \in \{0,1\}^{N\times N}$ represent the binary ground-truth mask indicating the annotated region (e.g., bounding box), where $M_{ij} = 1$ if patch (i, j) is inside the region, and 0 otherwise. Formally, AR is defined as $\mathrm{AR} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} A_{ij}\cdot M_{ij}}{\frac{\|A\|_1}{N^2}\cdot\|M\|_1}$, where $\|A\|_1 = \sum_{i=1}^{N}\sum_{j=1}^{N} A_{ij}$ and similarly for $\|M\|_1$.

New metrics to quantify model's attention map alignment. We note that AR only considers the amount of attention within the bounding box, ignoring how the attention is distributed. Particularly, a uniform distribution of attention within the bounding box region would be preferable, as questions in VGMED are specifically designed to require attention to entire bounding box regions. To take attention distribution into account, we propose to use the Kullback–Leibler (KL) and Jensen–Shannon (JS) divergence, which measure the difference between the attention map and bounding box by viewing them as two probability distributions.

Kullback–Leibler (KL) divergence. We compute the KL divergence between the normalized ground-truth mask $\hat{M}$ and the normalized attention map $\hat{A}$ as $D_{\mathrm{KL}}(\hat{M} \parallel \hat{A}) = \sum_{i=1}^{N}\sum_{j=1}^{N} \hat{M}_{ij} \log\left(\frac{\hat{M}_{ij}}{\hat{A}_{ij}}\right)$, where $\hat{A}_{ij} = A_{ij}/\|A\|_1$ and $\hat{M}_{ij} = M_{ij}/\|M\|_1$.

Jensen–Shannon (JS) divergence. To obtain a symmetric and bounded divergence metric, we compute the JS divergence as $D_{\mathrm{JS}}(\hat{M} \parallel \hat{A}) = \frac{1}{2}D_{\mathrm{KL}}(\hat{M} \parallel \hat{R}) + \frac{1}{2}D_{\mathrm{KL}}(\hat{A} \parallel \hat{R}), \quad \hat{R}_{ij} = \frac{1}{2}\left(\hat{M}_{ij} + \hat{A}_{ij}\right)$. The KL and JS divergences allow us to quantify not only whether the model attends to the correct region, but also how its attention is distributed within that region. A lower divergence indicates better alignment and more consistent attention over clinically relevant areas, offering a complementary perspective to AR.

2.3 Experimental Setups

We conduct our analysis on 8 SOTA medical MLLMs, including LLaVA-Med (Li et al., 2023a), LLaVA-Tri (Xie et al., 2025), HuatuoGPT-Vision-7B/34B (Chen et al., 2024a) (abbreviated as HuatuoGPT-V), VILA-M3-8B/13B (Nath et al., 2024), MedRegA (Wang et al., 2025), and a variant of HuatuoGPT-V—referred to as HuatuoGPT-V-Bio—where the original CLIP vision encoder is replaced with BiomedCLIP, a domain-specific encoder trained on biomedical data (see Supp H for details). To analyze attention behavior, we compute the mean attention map across all heads in each LLM layer. Inspired by Zhang et al. (2025a), we normalize the attention map using a reference attention map obtained from the generic prompt: *"Write a general description of the image."*. This normalization helps highlight regions relevant to the specific question.

We also include LLaVA-v1.5-7B (Liu et al., 2024a) results on *natural scene images*. As a general-domain MLLM, LLaVA demonstrates strong performance and exhibits good visual grounding, with attention distributions that align closely with ground-truth regions (Zhang et al., 2025a; Kang et al., 2025). This serves as a useful reference point for interpreting attention ratios, KL and JS divergence associated with effective visual grounding.

2.4 Empirical Analysis

Medical MLLMs exhibit inadequate visual grounding on medical images. We plot the attention ratio, KL divergence and JS divergence across all LLM layers for all models in Fig. 3. As shown

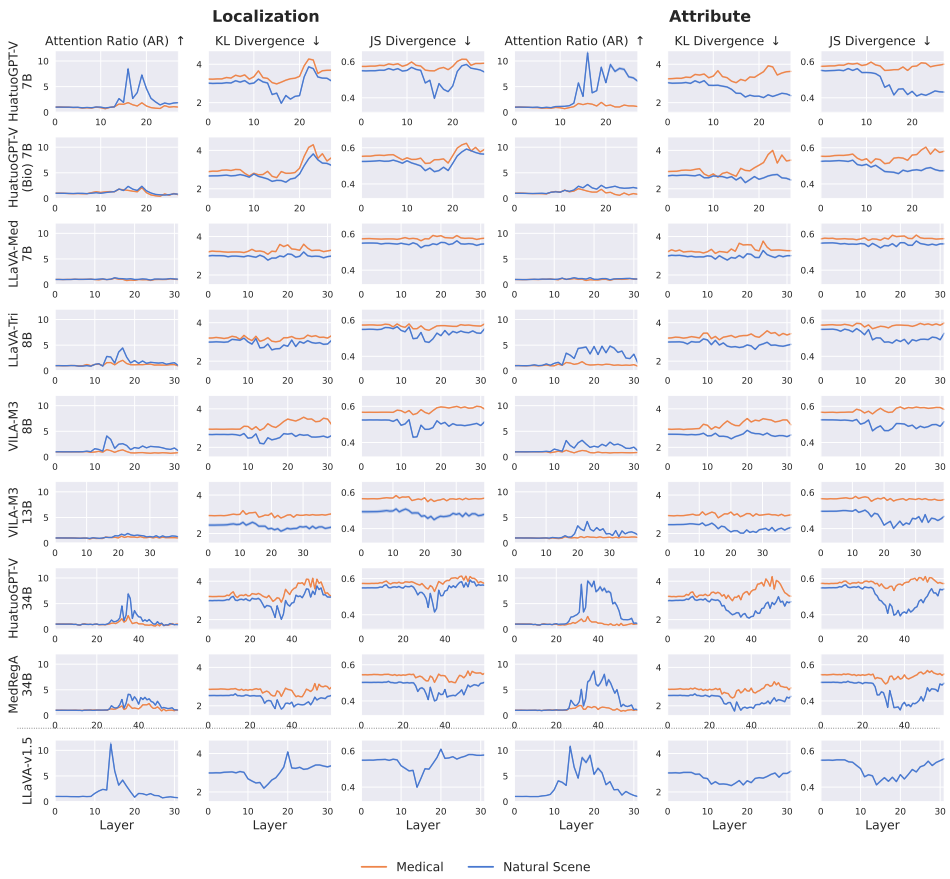

Figure 3: **Medical MLLMs demonstrate suboptimal visual grounding when applied to medical images.** Analysis using our proposed VGMED dataset—designed specifically to assess visual grounding in medical MLLMs—shows that all evaluated medical MLLMs exhibit substantial weaker alignment between their attention distributions and ground-truth annotations on medical images compared to natural scene images (from MS COCO). Additional comparison with general domain MLLM LLaVA-v1.5 on natural images (below the dashed line) further confirms that medical MLLMs consistently exhibit reduced alignment with annotated regions. **Best viewed in color and with zoom.**

in the figure, all evaluated medical MLLMs demonstrate weaker alignment between their attention distributions and ground-truth annotations when applied to medical images, compared to natural images. This is quantitatively and consistently supported by lower AR and higher values in our proposed KL and JS divergence metrics for measuring attention alignment. These trends persist across most network layers and are consistent for both attribute and localization tasks. Further comparison with LLaVA-v1.5 on natural images reinforces this observation: medical MLLMs show significantly lower alignment with annotated regions, as measured by AR, KL, and JS—highlighting deficiencies in visual grounding for medical image analysis.

For qualitative analysis, we visualize the attention map from the layer with the lowest KL divergence in Fig.1. Lower KL divergence reflects closer alignment between the model's attention distribution and the annotated regions, indicating that these layers are most relevant for visual grounding analysis. **Comprehensive qualitative analysis and visualization are included in Supp J.2**

3 Visual Grounding Refinement

Our analysis in Sec. 2 suggests that current medical MLLMs attend to clinically-relevant and irrelevant regions. In this section, we propose Visual Grounding Refinement (VGRefine), an inference-time method that enhances visual grounding in medical MLLMs by suppressing attention to clinically irrelevant regions. Specifically, as shown in Fig. 4 our method consists of two steps: 1) Attention Triage and 2) Attention Knockout.

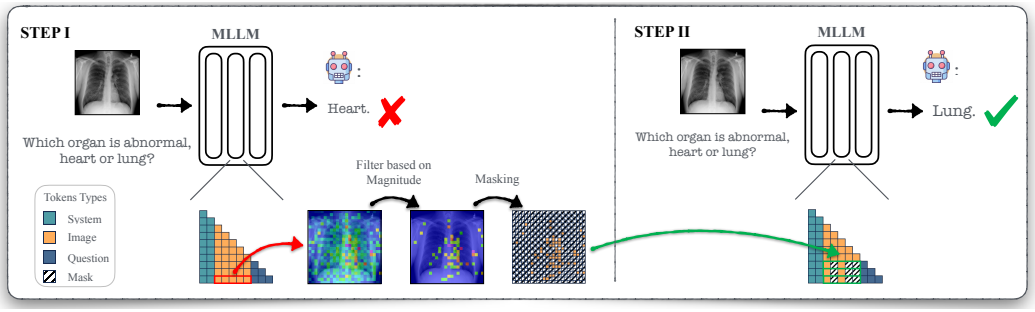

Figure 4: **Illustration of the proposed VGRefine method**: a two-step inference-time method to improve visual grounding in medical MLLMs. In **Step I (Attention Triage)**, we aggregate attention from the model's most visually sensitive heads and suppress low-confident attention, obtaining a binary mask. In **Step II (Attention Knockout)**, we use this mask to refine the model's attention distribution, improving its focus on relevant regions during inference. In the lower triangular attention matrix, each row represents the attention score of a query token to all key tokens.

Step I: Attention Triage — More Focusing on Clinically Relevant Regions. As illustrated in Fig. 1, medical MLLM's attention maps are often noisy—while they do attend to relevant areas, they also include a substantial focus on irrelevant regions, which diminishes interpretability and precision. To better focus on clinically meaningful regions, we move beyond layer-wise average attention and instead examine visual sensitivity at the head level across all layers. Following the same evaluation in Sec. 2.4, we identify the top K attention heads that most consistently align with visually relevant regions, using our proposed evaluation dataset (Sec. 2.1) and metric (Sec. 2.2). We then aggregate the attention maps from these selected heads with their average. We suppress low-activation regions based on magnitude of attention, as these are likely to represent irrelevant or noisy attention (see Supp. for further details and motivation). *This results in a sparse attention map with high-confidence.* We convert this filtered attention map into a binary mask $\mathcal{M} \in \{0,1\}^{N^2}$ by simply setting all non-zero entries to 1 and keeping the zeros unchanged.

Step II: Attention Knockout — Suppressing Irrelevant Visual Input. To enhance the visual grounding ability of medical MLLMs, we aim to guide the model's attention toward clinically relevant regions. Intuitively, improving focus on these regions can suppress distractions from irrelevant areas and yield more interpretable predictions. Similarly, recent advances in attention manipulation (Zhang et al., 2024; Geva et al., 2023; Zhang et al., 2025c) have shown that attending to redundant information potentially detriment to prediction as they distract the model's focus, they improve model behavior by preventing attention to uninformative tokens.

Inspired by this, we propose to knock out attention connections between question tokens and clinically irrelevant visual tokens. Specifically, we apply the binary mask $\mathcal{M}$ obtained in Step I to the attention weights $\alpha_q^{\ell,h}$, where $\alpha_q^{\ell,h}$ denotes the cross-attention from the qth question token to all visual tokens at layer ℓ and head h. We compute the masked attention as $\hat{\alpha}_q^{\ell,h} = \alpha_q^{\ell,h} \odot \mathcal{M}$, and use $\hat{\alpha}_q^{\ell,h}$ for the subsequent attention computation in model's forward pass. $\odot$ denotes element-wise multiplication. The masking operation explicitly restricts question tokens from receiving information from irrelevant visual regions at the selected layer. This modification encourages the model to attend selectively to meaningful regions, reducing distraction from irrelevant areas and therefore enhancing models' visual grounding capability.

4 Experiments

4.1 Evaluation Settings

Baselines. We compared two types of open-source models: (1) Medical MLLMs. We evaluated with the latest medical MLLMs, including Med-Flamingo (Moor et al., 2023), RadFM (Wu et al., 2023), LLaVA-Med-7B (Li et al., 2023a), LLaVA-Tri (Xie et al., 2025), MedPLIB (Huang et al., 2025), VILA-M3(Nath et al., 2024), HuatuoGPT-V (Chen et al., 2024a). (2) General MLLMs. We compared with two latest models pretrained on natural scene domain, LLaVA-v1.6-7B (Liu et al., 2024a) and Qwen-VL-Chat (Bai et al., 2023). We include the comparison of larger models in Supp.

Benchmarks. We follow the exact evaluation protocol of Chen et al. (2024a). Specifically, we adopt six benchmarks that are designed for biomedical MLLM evaluation, including VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021a), PathVQA (He et al., 2020), PMC-VQA (Zhang et al., 2023b), OmniMedVQA (Hu et al., 2024) (open-access split), and MMMU (Health & Medicine track) Yue et al. (2024). All evaluations were conducted in a zero-shot setting using question templates provided by LLaVA (details in Supp.).

VGRefine. We applied our inference-time method on HuatuoGPT-V (Chen et al., 2024a). All experiments are conducted using the same hyperparameters across benchmarks. Specifically, for Step I, we aggregate the attention maps from the top K heads with the highest alignment to visual relevant regions, as measured by KL divergence on our curated evaluation set built using COCO images. This setup prevents data leakage from medical evaluation benchmarks and demonstrates that our method generalizes from natural images to biomedical domains. Low-activation regions are suppressed based on a percentile threshold p over attention magnitude. We discuss the choice of K and p in Sec. 4.4. For Step II we apply the attention knockout only at layer $\ell = 16$ layer, which, according to our analysis in Fig. 3, demonstrates the most relevancy to visual grounding among all the layers.

4.2 Experimental Results

We follow exactly the evaluation setup of HuatuoGPT-V (Chen et al., 2024a) to ensure consistency across all benchmarks. Since the original papers of HuatuoGPT-V-7B (Chen et al., 2024a) and VILA-M3-8B (Nath et al., 2024) do not report results on certain benchmarks, we re-evaluate both models under the same zero-shot setting. For models with complete benchmark results available in their original publications—such as MedPLIB (Huang et al., 2025) and LLaVA-Tri (Xie et al., 2025)—we directly report the official numbers. For all other baselines, we use the results provided in the HuatuoGPT-V paper, as it adopts the same evaluation protocol.

It is important to note that some models include benchmark training sets during pretraining, making zero-shot evaluation unfair. Specifically, VILA-M3 (Nath et al., 2024) and MedPLIB (Huang et al., 2025) incorporate training data from VQA-RAD, SLAKE, PathVQA, and PMC-VQA, and thus are excluded from our zero-shot comparison on those datasets.

Medical VQA Benchmarks. Table 1 shows results on four standard medical VQA datasets. Here, we report the closed-ended question accuracy and a weighted average (Avg.) that scales by the number of samples in each benchmark (Additional results are in Supp.). Our inference-time method VGRefine applied to HuatuoGPT-V consistently improves its performance. We observe notable gains of +5.6% on VQA-RAD and +11.3% on PathVQA, with the overall average increasing from 65.3% to 68.4%, outperforming all baselines. These results underscore that enhanced visual grounding contributes to better performance on medical VQA tasks. On the MMMU benchmark (Table 2), VGRefine achieves the highest accuracy across all five sub-domains, increasing the overall average from 45.8% to 47.2%. This demonstrates that enhancing visual grounding at inference time also improves complex multimodal medical reasoning. As shown in Table 3, VGRefine improves performance across all eight imaging modalities, with significant boosts on CT (+7.5%), MRI (+6.4%), and X-Ray (+8.1%) on the OmniMedVQA benchmark. These results confirm the generalizability of our visual grounding refinement across diverse medical imaging tasks. Overall, our method raises average accuracy from 71.3% to 74.4%, demonstrating its robustness and generalizability across a wide range of modalities.

Table 1: Accuracy on medical VQA datasets. To align with the evaluation protocol with HuatuoGPT-V (Chen et al., 2024a), we specifically used the closed-ended subset for evaluation. Evaluation on other subsets in Supp.

Model	VQA-RAD	SLAKE	PathVQA	PMC-VQA	Avg.
Qwen-VL-Chat	47.0	56.0	55.1	36.6	48.9
LLaVA-v1.6-7B	52.6	57.9	47.9	35.5	48.5
Med-Flamingo	45.4	43.5	54.7	23.3	41.7
RadFM	50.6	34.6	38.7	25.9	37.5
LLaVA-Med-7B	51.4	48.6	56.8	24.7	45.4
LLaVA-Tri	59.8	43.4	59.0	-	-
HuatuoGPT-V-7B	67.4	76.5	60.7	53.9	65.3
VGRefine (Ours)	**71.2**	**76.9**	**67.6**	**56.2**	**68.4**

Table 2: Accuracy on MMMU Health & Medicine benchmark. **BMS**, **CM**, **DLM**, **P**, **PH** denote Basic Medical Science, Clinical Medicine, Diagnostics & Laboratory Medicine, Pharmacy, Public Health respectively.

Model	BMS	CM	DLM	P	PH	Avg.
Qwen-VL-Chat	36.5	31.7	32.7	28.4	34.6	32.7
LLaVA-v1.6-7B	40.5	36.9	32.1	32.3	26.9	33.1
Med-Flamingo	29.6	28.1	24.8	25.3	31.2	28.3
RadFM	27.5	26.8	25.8	24.7	29.1	27.0
LLaVA-Med-7B	39.9	39.1	34.6	37.4	34.0	36.9
LLaVA-Tri	37.1	-	27.8	-	-	-
VILA-M3-8B	39.3	39.7	34.0	32.1	28.7	34.0
HuatuoGPT-V-7B	58.9	57.2	43.8	37.2	38.3	45.8
VGRefine (Ours)	**59.5**	**59.1**	**45.7**	**38.6**	**39.3**	**47.2**

Table 3: The accuracy of OmniMedVQA within different modalities. Specifically, **FP** denotes *Fundus Photography*, **MRI** denotes *Magnetic Resonance Imaging*, **OCT** denotes *Optical Coherence Tomography*, **Der** denotes *Dermoscopy*, **Mic** denotes *Microscopy Images*, **US** denotes *Ultrasound*.

Model	CT	FP	MRI	OCT	Der	Mic	X-Ray	US	Avg.
Qwen-VL-Chat (Bai et al., 2023)	51.5	45.4	43.9	54.0	55.4	49.5	63.1	33.5	49.5
LLaVA-v1.6-7B (Liu et al., 2024a)	40.1	39.5	54.8	58.4	54.0	48.8	53.3	47.9	49.6
Med-Flamingo (Moor et al., 2023)	34.6	33.3	27.5	26.0	28.3	28.1	30.1	33.2	30.2
RadFM (Wu et al., 2023)	33.3	35.0	22.0	31.3	36.3	28.0	31.5	26.1	30.5
LLaVA-Med-7B (Li et al., 2023a)	25.3	48.4	35.9	42.1	45.2	44.0	31.7	34.4	35.8
VILA-M3-8B (Nath et al., 2024)	60.2	35.7	51.5	56.9	51.5	51.7	65.4	46.1	53.0
MedPLIB (Huang et al., 2025)	62.7	65.0	67.0	75.1	51.5	64.4	60.3	38.8	60.6
HuatuoGPT-V-7B (Chen et al., 2024a)	62.6	80.3	67.7	86.2	**71.7**	74.2	74.2	**79.7**	71.3
VGRefine (Ours)	**67.3**	**82.4**	**72.0**	**86.9**	**71.7**	**74.9**	**80.2**	79.5	**74.4**

Table 4: Ablation study on the choice of top K heads and p precentile of magnitude-based filtering.

K	VQA-RAD	SLAKE	PathVQA	PMC-VQA	Avg.	p (%)	VQA-RAD	SLAKE	PathVQA	PMC-VQA	Avg.
1	68.62	75.81	64.85	53.65	68.28	30	70.78	76.84	67.55	55.70	68.22
2	69.78	76.63	63.58	53.90	68.21	40	70.70	76.66	67.28	56.00	68.12
5	70.09	76.56	66.52	54.30	68.08	50	**71.24**	**76.88**	67.61	**56.20**	**68.42**
8	70.70	76.52	66.64	55.60	68.12	60	70.70	76.66	67.28	55.65	68.05
10	70.86	76.81	**67.67**	56.05	68.34	70	70.47	76.66	67.61	55.80	68.17
15	70.78	76.84	67.43	55.40	68.11	80	70.78	76.73	67.55	55.80	68.21
20	**71.24**	**76.88**	67.61	**56.20**	**68.42**	90	70.55	76.34	**68.11**	55.50	68.20

4.3 Human Evaluation: VGRefine improves trustworthiness

We conducted a blinded study with five experienced clinicians using 20 medical VQA cases from VGMED. Each case presented two attention maps: one from the baseline model and one from the same model after applying VGRefine. The source of each attention map was not disclosed, and their order was randomized. Clinicians were asked which map appeared more clinically reasonable and trustworthy. VGRefine was preferred in 76% of cases, with feedback noting improved focus and reduced noise. These results suggest that VGRefine enhances clinician trust by producing more interpretable visual. See human evaluation details in Supp J.1.

4.4 Ablation Studies

Table 4 presents ablations on the number of top attention heads K and the percentile threshold p used for magnitude-based filtering. Performance improves consistently as K increases, with the best average accuracy (68.42%) achieved at $K = 20$, indicating that aggregating more heads helps capture richer grounding signals. For the percentile p, the model remains stable across values, with optimal performance also at $p = 50\%$, confirming the effectiveness of moderate filtering in removing noisy regions without discarding relevant information.

5 Conclusion

In this work, we presented the first systematic analysis of visual grounding in medical MLLMs. Using our clinically guided VGMED dataset and newly introduced metrics, we showed across 8 SOTA medical MLLMs frequent failures in grounding predictions in clinically relevant regions. This failure mode persisted even in recent medical MLLMs and contributed to their underperformance in zero-shot medical image understanding. To address this, we proposed VGRefine, an inference-time attention refinement method to improve medical MLLMs' visual grounding. Across 6 diverse Med-VQA benchmarks, comprising over 110K VQA samples from 8 imaging modalities, VGRefine consistently achieves SOTA performance. We remark that improvements using VGRefine are achieved without retraining or introducing any new medical knowledge. If visual grounding were not a limiting factor, such consistent gains would not occur. Therefore, VGRefine results further support that visual grounding deficiency is a general, widespread issue. Overall, our proposed VGMED helps uncover and confirm inadequate visual grounding, while VGRefine experiments demonstrate its broad prevalence and generalization across different modalities and clinical scenarios. Our findings underscored the importance of grounding-aware analysis to achieve more reliable and generalizable medical MLLMs. **Additional experiments, limitation and ethical consideration are included in Supp.**

REFERENCES

W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, February 2020. doi: 10.1016/j.dib.2019.104863. Retrieved: September 25, 2025.

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C. Kitamura, Sarthak Pati, Luciano M. Prevedello, Jeffrey D. Rudie, Chiharu Sako, Russell T. Shinohara, Timothy Bergquist, Rong Chai, James Eddy, Julia Elliott, Walter Reade, Thomas Schaffter, Thomas Yu, Jiaxin Zheng, Ahmed W. Moawad, Luiz Otavio Coelho, Olivia McDonnell, Elka Miller, Fanny E. Moron, Mark C. Oswood, Robert Y. Shih, Loizos Siakallis, Yulia Bronstein, James R. Mason, Anthony F. Miller, Gagandeep Choudhary, Aanchal Agarwal, Cristina H. Besada, Jamal J. Derakhshan, Mariana C. Diogo, Daniel D. Do-Dai, Luciano Farage, John L. Go, Mohiuddin Hadi, Virginia B. Hill, Michael Iv, David Joyner, Christie Lincoln, Eyal Lotan, Asako Miyakoshi, Mariana Sanchez-Montano, Jaya Nath, Xuan V. Nguyen, Manal Nicolas-Jilwan, Johanna Ortiz Jimenez, Kerem Ozturk, Bojan D. Petrovic, Chintan Shah, Lubdha M. Shah, Manas Sharma, Onur Simsek, Achint K. Singh, Salil Soman, Volodymyr Statsevych, Brent D. Weinberg, Robert J. Young, Ichiro Ikuta, Amit K. Agarwal, Sword C. Cambron, Richard Silbergleit, Alexandru Dusoi, Alida A. Postma, Laurent Letourneau-Guillon, Gloria J. Guzman Perez-Carrillo, Atin Saha, Neetu Soni, Greg Zaharchuk, Vahe M. Zohrabian, Yingming Chen, Milos M. Cekic, Akm Rahman, Juan E. Small, Varun Sethi, Christos Davatzikos, John Mongan, Christopher Hess, Soonmee Cha, Javier Villanueva-Meyer, John B. Freymann, Justin S. Kirby, Benedikt Wiestler, Priscila Crivellaro, Rivka R. Colen, Aikaterini Kotrotsou, Daniel Marcus, Mikhail Milchenko, Arash Nazeri, Hassan Fathallah-Shaykh, Roland Wiest, Andras Jakab, Marc-Andre Weber, Abhishek Mahajan, Bjoern Menze, Adam E. Flanders, and Spyridon Bakas. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification, 2021. URL https://arxiv.org/abs/2107.02314.

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1): 1–13, 2017.

Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111, 2015.

Aditya Bharatha, Masanori Hirose, Nobuhiko Hata, Simon K. Warfield, Matthieu Ferrant, Kelly H. Zou, Eduardo Suarez-Santana, Juan Ruiz-Alzola, Anthony D'Amico, Robert A. Cormack, Ron Kikinis, Ferenc A. Jolesz, and Clare M. C. Tempany. Evaluation of three-dimensional finite element-based deformable registration of pre- and intraoperative prostate imaging. *Medical Physics*, 28(12):2551–2560, 2001. doi: https://doi.org/10.1118/1.1414009. URL https://aapm.onlinelibrary.wiley.com/doi/abs/10.1118/1.1414009.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, and Benyou Wang. Towards injecting medical visual knowledge into multimodal LLMs at scale. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7346–7370, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.418. URL https://aclanthology.org/2024.emnlp-main.418/.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler,

and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 19–35, Cham, 2025a. Springer Nature Switzerland. ISBN 978-3-031-73004-7.

Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. Why is spatial reasoning hard for vlms? an attention mechanism perspective on focus areas. In *Forty-second International Conference on Machine Learning*, 2025b.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024b.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic), 2019. URL https://arxiv.org/abs/1902.03368.

Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 168–172, 2018a. doi: 10.1109/ISBI.2018.8363547.

Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 168–172, 2018b. doi: 10.1109/ISBI.2018.8363547.

Hejie Cui, Lingjun Mao, Xin Liang, Jieyu Zhang, Hui Ren, Quanzheng Li, Xiang Li, and Carl Yang. Biomedical visual instruction tuning with clinician preference alignment. 6 2024. doi: 10.48550/arXiv.2406.13173. URL https://arxiv.org/abs/2406.13173v3.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=vvoWPYqZJA.

OpenAI et al. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Huihui Fang, Fei Li, Huazhu Fu, Xu Sun, Xingxing Cao, Fengbin Lin, Jaemin Son, Sunho Kim, Gwenole Quellec, Sarah Matta, Sharath M. Shankaranarayana, Yi-Ting Chen, Chuen-Heng Wang, Nisarg A. Shah, Chia-Yen Lee, Chih-Chung Hsu, Hai Xie, Baiying Lei, Ujjwal Baid, Shubham Innani, Kang Dang, Wenxiu Shi, Ravi Kamble, Nitin Singhal, Ching-Wei Wang, Shih-Chang Lo, José Ignacio Orlando, Hrvoje Bogunović, Xiulan Zhang, and Yanwu Xu. Adam challenge: Detecting age-related macular degeneration from fundus images. *IEEE Transactions on Medical Imaging*, 41(10):2828–2847, 2022. doi: 10.1109/TMI.2022.3172773.

Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision*, pp. 3429–3437, 2017.

Huazhu Fu, Fei Li, José Ignacio Orlando, Hrvoje Bogunovic, Xu Sun, Jingan Liao, Yanwu Xu, Shaochong Zhang, and Xiulan Zhang. ichallenge-palm: Pathologic myopia challenge, 07 2019. URL https://cir.nii.ac.jp/crid/1880020692683121792.

Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. Interpreting clip's image representation via text-based decomposition. In *The Twelfth International Conference on Learning Representations*, 2024.

Sergios Gatidis, Tobias Hepp, Marcel Früh, Christian La Fougère, Konstantin Nikolaou, Christina Pfannenberg, Bernhard Schölkopf, Thomas Küstner, Clemens Cyran, and Daniel Rubin. A whole-body fdg-pet/ct dataset with manually annotated tumor lesions. *Scientific Data*, 9(1):601, 2022.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235, 2023.

Michal Golovanevsky, William Rudman, Vedant Palit, Ritambhara Singh, and Carsten Eickhoff. What do vlms notice? a mechanistic interpretability pipeline for noise-free text-image corruption and evaluation. *CoRR*, abs/2406.16320, 2024. URL https://doi.org/10.48550/arXiv.2406.16320.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering, 2020. URL https://arxiv.org/abs/2003.10286.

Nicholas Heller, Niranjan Sathianathen, Arveen Kalapara, Edward Walczak, Keenan Moore, Heather Kaluzniak, Joel Rosenberg, Paul Blake, Zachary Rengel, Makinna Oestreich, Joshua Dean, Michael Tradewell, Aneri Shah, Resha Tejpaul, Zachary Edgerton, Matthew Peterson, Shaneabbas Raza, Subodh Regmi, Nikolaos Papanikolopoulos, and Christopher Weight. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes, 2020. URL https://arxiv.org/abs/1904.00445.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 32, 2019.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22170–22183, June 2024.

Rui Huang, Zijie Chen, Yuanyuan Chen, Hongsheng Li, et al. StructSeg2019 Grand Challenge Dataset. https://structseg2019.grand-challenge.org/Dataset/, 2019. Retrieved: September 25, 2025.

Xiaoshuang Huang, Lingdong Shen, Jia Liu, Fangxin Shang, Hongxiang Li, Haifeng Huang, and Yehui Yang. Towards a multimodal large language model with pixel-level insight for biomedicine. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(4):3779–3787, Apr. 2025. doi: 10.1609/aaai.v39i4.32394. URL https://ojs.aaai.org/index.php/AAAI/article/view/32394.

J Igelsias, M Styner, T Langerak, B Landman, Z Xu, and A Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, 2015.

Aya Abdelsalam Ismail, Hector Corrada Bravo, and Soheil Feizi. Improving deep learning interpretability by saliency guided training. *Advances in Neural Information Processing Systems*, 34: 26726–26739, 2021.

Daniel P Jeong, Saurabh Garg, Zachary Chase Lipton, and Michael Oberst. Medical adaptation of large language and vision-language models: Are we making progress? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 12143–12170, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.677. URL https://aclanthology.org/2024.emnlp-main.677/.

Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *International conference on multimedia modeling*, pp. 451–462. Springer, 2019.

Debesh Jha, Nikhil Kumar Tomar, Vanshali Sharma, Quoc-Huy Trinh, Koushik Biswas, Hongyi Pan, Ritika K. Jha, Gorkem Durak, Alexander Hann, Jonas Varkey, Hang Viet Dao, Long Van Dao, Binh Phuc Nguyen, Nikolaos Papachrysos, Brandon Rieders, Peter Thelin Schmidt, Enrik

Geissler, Tyler Berzin, Pål Halvorsen, Michael A. Riegler, Thomas de Lange, and Ulas Bagci. Polypdb: A curated multi-center dataset for development of ai algorithms in colonoscopy, 2025. URL https://arxiv.org/abs/2409.00045.

Yuanfeng Ji, Haotian Bai, Chongjian GE, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, and Ping Luo. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 36722–36732. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/ee604e1bedbd069d9fc9328b7b9584be-Paper-Datasets_and_Benchmarks.pdf.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25004–25014, June 2025.

Omri Kaduri, Shai Bagon, and Tali Dekel. What's in the image? a deep-dive into the vision of vision language models, 2024. URL https://arxiv.org/abs/2411.17491.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. Your large vision-language model only needs a few attention heads for visual grounding, 2025. URL https://arxiv.org/abs/2503.06287.

A. Emre Kavur, N. Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, Bora Baydar, Dmitry Lachinov, Shuo Han, Josef Pauli, Fabian Isensee, Matthias Perkonigg, Rachana Sathish, Ronnie Rajan, Debdoot Sheet, Gurbandurdy Dovletov, Oliver Speck, Andreas Nürnberger, Klaus H. Maier-Hein, Gözde Bozdağı Akar, Gözde Ünal, Oğuz Dicle, and M. Alper Selver. Chaos challenge - combined (ct-mr) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, 2021. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2020.101950. URL https://www.sciencedirect.com/science/article/pii/S1361841520303145.

Jason J. Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5:1–10, 11 2018.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 28541–28564. Curran Associates, Inc., 2023a. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/5abcdf8ecdcacba028c6662789194572-Paper-Datasets_and_Benchmarks.pdf.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 23–29 Jul 2023b. URL https://proceedings.mlr.press/v202/li23q.html.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

Geert Litjens, Robert Toth, Wendy van de Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, Robin Strand, Filip Malmberg, Yangming Ou, Christos Davatzikos, Matthias Kirschner, Florian Jung, Jing Yuan, Wu Qiu, Qinquan Gao, Philip "Eddie" Edwards, Bianca Maan, Ferdinand van der Heijden, Soumya Ghose, Jhimli

Mitra, Jason Dowling, Dean Barratt, Henkjan Huisman, and Anant Madabhushi. Evaluation of prostate segmentation algorithms for mri: The promise12 challenge. *Medical Image Analysis*, 18(2): 359–373, 2014. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2013.12.002. URL https://www.sciencedirect.com/science/article/pii/S1361841513001734.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1650–1654, 2021a. doi: 10.1109/ISBI48211.2021.9434010.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=w0H2xGHlkw.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26296–26306, June 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024b. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Pengbo Liu, Hu Han, Yuanqi Du, Heqin Zhu, Yinhao Li, Feng Gu, Honghu Xiao, Jun Li, Chunpeng Zhao, Li Xiao, et al. Deep learning to segment pelvic bones: large-scale ct datasets and baseline models. *International Journal of Computer Assisted Radiology and Surgery*, 16(5):749–756, 2021b.

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, pp. 125–140. Springer, 2024c.

Wentao Liu, Qianjun Pan, Yi Zhang, Zhuo Liu, Ji Wu, Jie Zhou, Aimin Zhou, Qin Chen, Bo Jiang, and Liang He. Cmm-math: A chinese multimodal math dataset to evaluate and enhance the mathematics reasoning of large multimodal models. *CoRR*, abs/2409.02834, 2024d. URL https://doi.org/10.48550/arXiv.2409.02834.

Jiaying Lu, Jinmeng Rao, Kezhen Chen, Xiaoyuan Guo, Yawen Zhang, Baochen Sun, Carl Yang, and Jie Yang. Evaluation and enhancement of semantic grounding in large vision-language models. In *AAAI-24 Workshop on Responsible Language Models*, 2024. URL https://arxiv.org/abs/2309.04041.

Xiangde Luo, Wenjun Liao, Jianghong Xiao, Jieneng Chen, Tao Song, Xiaofan Zhang, Kang Li, Dimitris N. Metaxas, Guotai Wang, and Shaoting Zhang. Word: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from ct image. *Medical Image Analysis*, 82:102642, 2022. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2022.102642. URL https://www.sciencedirect.com/science/article/pii/S1361841522002705.

Holger Lyre. Understanding ai: Semantic grounding in large language models, 2024. URL https://arxiv.org/abs/2402.10992.

Jun Ma, Yao Zhang, Song Gu, Xingle An, Zhihe Wang, Cheng Ge, Congcong Wang, Fan Zhang, Yu Wang, Yinan Xu, Shuiping Gou, Franz Thaler, Christian Payer, Darko Štern, Edward G.A. Henderson, Dónal M. McSweeney, Andrew Green, Price Jackson, Lachlan McIntosh, Quoc-Cuong Nguyen, Abdul Qayyum, Pierre-Henri Conze, Ziyan Huang, Ziqi Zhou, Deng-Ping Fan, Huan Xiong, Guoqiang Dong, Qiongjie Zhu, Jian He, and Xiaoping Yang. Fast and low-gpu-memory abdomen ct organ segmentation: The flare challenge. *Medical Image Analysis*, 82:102616, 2022a. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2022.102616. URL https://www.sciencedirect.com/science/article/pii/S1361841522002444.

Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, Shucheng Cao, Qi Zhang, Shangqing Liu, Yunpeng Wang, Yuhui Li, Jian He, and Xiaoping Yang. Abdomenct-1k: Is abdominal organ segmentation a solved problem? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6695–6714, 2022b. doi: 10.1109/TPAMI.2021.3100536.

Jun Ma, Yao Zhang, Song Gu, Cheng Ge, Shihao Mae, Adamo Young, Cheng Zhu, Xin Yang, Kangkang Meng, Ziyan Huang, et al. Unleashing the strengths of unlabelled data in deep learning-assisted pan-cancer abdominal organ quantification: the flare22 challenge. *The Lancet Digital Health*, 6(11):e815–e826, 2024.

Oskar Maier, Bjoern H. Menze, Janina von der Gablentz, Levin Häni, Mattias P. Heinrich, Matthias Liebrand, Stefan Winzeck, Abdul Basit, Paul Bentley, Liang Chen, Daan Christiaens, Francis Dutil, Karl Egger, Chaolu Feng, Ben Glocker, Michael Götz, Tom Haeck, Hanna-Leena Halme, Mohammad Havaei, Khan M. Iftekharuddin, Pierre-Marc Jodoin, Konstantinos Kamnitsas, Elias Kellner, Antti Korvenoja, Hugo Larochelle, Christian Ledig, Jia-Hong Lee, Frederik Maes, Qaiser Mahmood, Klaus H. Maier-Hein, Richard McKinley, John Muschelli, Chris Pal, Linmin Pei, Janaki Raman Rangarajan, Syed M.S. Reza, David Robben, Daniel Rueckert, Eero Salli, Paul Suetens, Ching-Wei Wang, Matthias Wilms, Jan S. Kirschke, Ulrike M. Krämer, Thomas F. Münte, Peter Schramm, Roland Wiest, Heinz Handels, and Mauricio Reyes. Isles 2015 - a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral mri. *Medical Image Analysis*, 35:250–269, 2017. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2016.07.009. URL https://www.sciencedirect.com/science/article/pii/S1361841516301268.

Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, Levente Lanczi, Elizabeth Gerstner, Marc-André Weber, Tal Arbel, Brian B. Avants, Nicholas Ayache, Patricia Buendia, D. Louis Collins, Nicolas Cordier, Jason J. Corso, Antonio Criminisi, Tilak Das, Hervé Delingette, Çağatay Demiralp, Christopher R. Durst, Michel Dojat, Senan Doyle, Joana Festa, Florence Forbes, Ezequiel Geremia, Ben Glocker, Polina Golland, Xiaotao Guo, Andac Hamamci, Khan M. Iftekharuddin, Raj Jena, Nigel M. John, Ender Konukoglu, Danial Lashkari, José António Mariz, Raphael Meier, Sérgio Pereira, Doina Precup, Stephen J. Price, Tammy Riklin Raviv, Syed M. S. Reza, Michael Ryan, Duygu Sarikaya, Lawrence Schwartz, Hoo-Chang Shin, Jamie Shotton, Carlos A. Silva, Nuno Sousa, Nagesh K. Subbanna, Gabor Szekely, Thomas J. Taylor, Owen M. Thomas, Nicholas J. Tustison, Gozde Unal, Flor Vasseur, Max Wintermark, Dong Hye Ye, Liang Zhao, Binsheng Zhao, Darko Zikic, Marcel Prastawa, Mauricio Reyes, and Koen Van Leemput. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2015. doi: 10.1109/TMI.2014.2377694.

Anna Montoya, Hasnin, kaggle446, shirzad, Will Cukierski, and yffud. Ultrasound Nerve Segmentation. https://kaggle.com/competitions/ultrasound-nerve-segmentation, 2016. Retrieved: September 25, 2025.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In Stefan Hegselmann, Antonio Parziale, Divya Shanmugam, Shengpu Tang, Mercy Nyamewaa Asiedu, Serina Chang, Tom Hartvigsen, and Harvineet Singh (eds.), *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pp. 353–367. PMLR, 10 Dec 2023. URL https://proceedings.mlr.press/v225/moor23a.html.

Vishwesh Nath, Wenqi Li, Dong Yang, Andriy Myronenko, Mingxin Zheng, Yao Lu, Zhijian Liu, Hongxu Yin, Yee Man Law, Yucheng Tang, et al. Vila-m3: Enhancing vision-language models with medical expert knowledge. *arXiv preprint arXiv:2411.12915*, 2024.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Danielle F Pace, Adrian V Dalca, Tal Geva, Andrew J Powell, Mehdi H Moghari, and Polina Golland. Interactive whole-heart segmentation in congenital heart disease. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 80–88. Springer, 2015.

Vedant Palit, Rohan Pandey, Aryaman Arora, and Paul Pu Liang. Towards vision-language mechanistic interpretability: A causal tracing tool for blip. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 2856–2861, October 2023.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas de Bel, Moira S.N. Berens, Cas van den Bogaard, Piergiorgio Cerello, Hao Chen, Qi Dou, Maria Evelina Fantacci, Bram Geurts, Robbert van der Gugten, Pheng Ann Heng, Bart Jansen, Michael M.J. de Kaste, Valentin Kotov, Jack Yu-Hung Lin, Jeroen T.M.C. Manders, Alexander Sóñora-Mengana, Juan Carlos García-Naranjo, Evgenia Papavasileiou, Mathias Prokop, Marco Saletta, Cornelia M Schaefer-Prokop, Ernst T. Scholten, Luuk Scholten, Miranda M. Snoeren, Ernesto Lopez Torres, Jef Vandemeulebroucke, Nicole Walasek, Guido C.A. Zuidhof, Bram van Ginneken, and Colin Jacobs. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The luna16 challenge. *Medical Image Analysis*, 42: 1–13, 2017. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2017.06.015. URL https://www.sciencedirect.com/science/article/pii/S1361841517301020.

Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models. *CoRR*, abs/2406.17294, 2024. URL https://doi.org/10.48550/arXiv.2406.17294.

Aliyun Tianchi. Chest image dataset for pneumothorax segmentation. https://tianchi.aliyun.com/dataset/83075, 2020. Accessed: September 25, 2025.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418*, 2019.

Lehan Wang, Haonan Wang, Honglong Yang, Jiaji Mao, Zehong Yang, Jun Shen, and Xiaomeng Li. Interpretable bilingual multimodal large language model for diverse biomedical tasks. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=YuHQTo6G9S.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024. URL https://arxiv.org/abs/2409.12191.

Jakob Wasserthal, Hanns-Christian Breit, Manfred T. Meyer, Maurice Pradella, Daniel Hinck, Alexander W. Sauter, Tobias Heye, Daniel T. Boll, Joshy Cyriac, Shan Yang, Michael Bach, and Martin Segeroth. Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5):e230024, 2023. doi: 10.1148/ryai.230024. URL https://doi.org/10.1148/ryai.230024.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology. *CoRR*, abs/2308.02463, 2023. URL https://doi.org/10.48550/arXiv.2308.02463.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. NExT-GPT: Any-to-any multimodal LLM. In *Proceedings of the International Conference on Machine Learning*, pp. 53366–53397, 2024.

Linhui Xiao, Xiaoshan Yang, Xiangyuan Lan, Yaowei Wang, and Changsheng Xu. Towards visual grounding: A survey, 2024. URL https://arxiv.org/abs/2412.20206.

Yunfei Xie, Ce Zhou, Lang Gao, Juncheng Wu, Xianhang Li, Hong-Yu Zhou, Sheng Liu, Lei Xing, James Zou, Cihang Xie, and Yuyin Zhou. Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=IwgmgidYPS.

Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. Mitigating hallucination in large vision-language models via modular attribution and intervention. In *Adaptive Foundation Models: Evolving AI for Personalized and Efficient Learning*, 2025.

Qinan Yu, Jack Merullo, and Ellie Pavlick. Characterizing mechanisms for factual recall in language models. In *The Conference on Empirical Methods in Natural Language Processing*, 2023.

Zeping Yu and Sophia Ananiadou. Understanding multimodal llms: the mechanistic interpretability of llava in visual question answering, 2025. URL https://arxiv.org/abs/2411.10950.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9556–9567, June 2024.

Anna Zawacki, Carol Wu, George Shih, Julia Elliott, Mikhail Fomitchev, Mohannad Hussain, Paras Lakhani, Phil Culliton, and Shunxing Bao. Siim-acr pneumothorax segmentation. https://kaggle.com/competitions/siim-acr-pneumothorax-segmentation, 2019. Kaggle.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. Mllms know where to look: Training-free perception of small visual details with multimodal llms. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=DgaY5mDdmThttps://github.com/saccharomycetes/mllms_know.

Minghui Zhang, Yangqian Wu, Hanxiao Zhang, Yulei Qin, Hao Zheng, Wen Tang, Corey Arnold, Chenhao Pei, Pengxin Yu, Yang Nan, Guang Yang, Simon Walsh, Dominic C. Marshall, Matthieu Komorowski, Puyang Wang, Dazhou Guo, Dakai Jin, Ya'nan Wu, Shuiqing Zhao, Runsheng Chang, Boyu Zhang, Xing Lu, Abdul Qayyum, Moona Mazher, Qi Su, Yonghuang Wu, Ying'ao Liu, Yufei Zhu, Jiancheng Yang, Ashkan Pakzad, Bojidar Rangelov, Raul San Jose Estepar, Carlos Cano Espinosa, Jiayuan Sun, Guang-Zhong Yang, and Yun Gu. Multi-site, multi-domain airway tree modeling. *Medical Image Analysis*, 90:102957, 2023a. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2023.102957. URL https://www.sciencedirect.com/science/article/pii/S1361841523002177.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1):AIoa2400640, 2025b. doi: 10.1056/AIoa2400640. URL https://ai.nejm.org/doi/full/10.1056/AIoa2400640.

Xiaofeng Zhang, Yihao Quan, Chen Shen, Xiaosong Yuan, Shaotian Yan, Liang Xie, Wenxiao Wang, Chaochen Gu, Hao Tang, and Jieping Ye. From redundancy to relevance: Enhancing explainability in multimodal large language models. *Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*, 2025c.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *CoRR*, abs/2305.10415, 2023b. URL https://doi.org/10.48550/arXiv.2305.10415.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models. *arXiv preprint arXiv:2411.18620*, 2024.

Zhongchen Zhao, Huai Chen, and Lisheng Wang. A coarse-to-fine framework fornbsp;thenbsp;2021 kidney andnbsp;kidney tumor segmentation challenge. In *Kidney and Kidney Tumor Segmentation: MICCAI 2021 Challenge, KiTS 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings*, pp. 53–58, Berlin, Heidelberg, 2021. Springer-Verlag. ISBN 978-3-030-98384-0. doi: 10.1007/978-3-030-98385-7_8. URL https://doi.org/10.1007/978-3-030-98385-7_8.

Wenwen Zhuang, Xin Huang, Xiantao Zhang, and Jin Zeng. Math-puma: Progressive upward multimodal alignment to enhance mathematical reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24):26183–26191, Apr. 2025. doi: 10.1609/aaai.v39i24.34815. URL https://ojs.aaai.org/index.php/AAAI/article/view/34815.

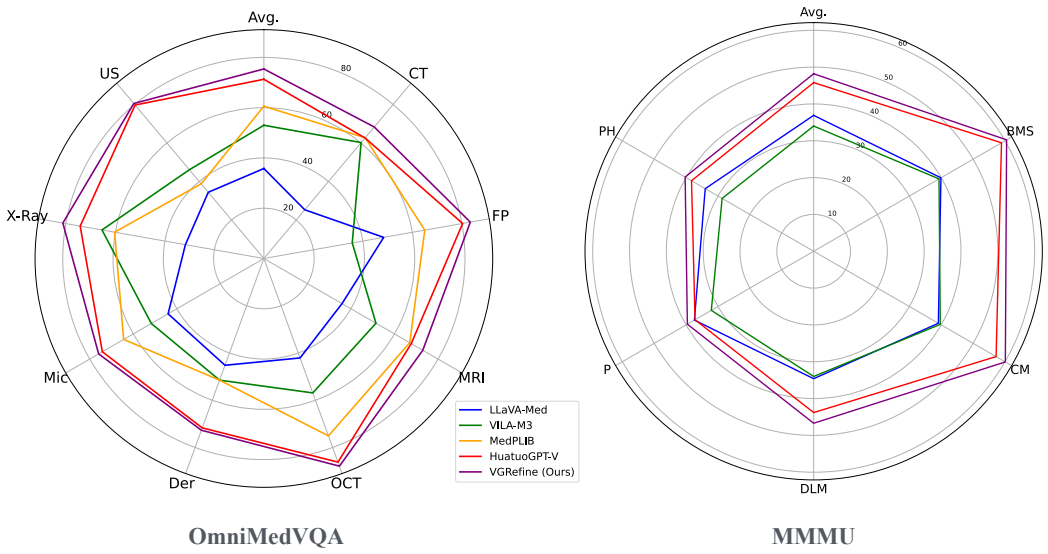

Figure A.1: Our proposed inference-time method VGRefine achieve state-of-the-art performance on OmniMedVQA (Hu et al., 2024) and MMMU (Health & Medicine track) (Yue et al., 2024). Many existing medical MLLMs remain to underperform on medical VQA tasks in the zero-shot setting as shown in this figure, but there is a lack of systematic study to understand the reasons. Compared to existing medical MLLMs, our proposed VGRefine demonstrates consistently stronger zero-shot performance across all modalities and sub-domains, highlighting its effectiveness in mitigating the issue of inadequate visual grounding as revealed in our study.

Appendix Overview

In this supplementary material, we provide additional experiments, ablation studies, and reproducibility details to support our findings. These sections are not included in the main paper due to space constraints.

Please find the following anonymous link for code and other resources: `https://anonymous.4open.science/r/Medical_MLLMs_Fail-8120/`.

Contents

A More Discussion on Related Work

Medical Multimodal Large Language Models (MLLMs). Recent advances in medical multimodal large language models (MLLMs) have focused on leveraging image-text pairs from sources like PubMed central (Zhang et al., 2023b; Moor et al., 2023; Chen et al., 2024a; Li et al., 2023a) and medical textbooks (Moor et al., 2023) to enable generative VQA and medical reasoning. Models such as LLaVA-Med (Li et al., 2023a), MedVInT (Zhang et al., 2023b), Med-Flamingo (Moor et al., 2023), HuatuoGPT-Vision (Chen et al., 2024a), and BioMed-VITAL (Cui et al., 2024) introduce GPT-4 (et al., 2024) generated instruction-following datasets and expert-validated responses to improve medical VQA performance. More recent studies have begun to explore different ideas to improve region awareness in biomedical MLLMs: explicit fine-tuning with additional supervision, such as annotated bounding boxes (Wang et al., 2025; Xie et al., 2025) or segmentation masks (Jeong et al., 2024), along with architectural modifications to support spatial reasoning. For instance, models like MedRegA (Wang et al., 2025) and LLaVA-Tri (Xie et al., 2025) rely on additional datasets. Other recent models focus on scale or domain expertise. VILA-M3 (Nath et al., 2024), for instance, incorporates domain-specific expert models during training, arguing that generic Vision–Language Models (VLMs) lack the fine-grained expertise needed for healthcare. Given their dependence on task-specific fine-tuning (Nath et al., 2024; Xie et al., 2025) and sub-optimal generalization in zero-shot settings (Li et al., 2023a), *it remains unclear whether current medical MLLMs ground their predictions in meaningful visual evidence within medical images.* To our knowledge, no prior work has conducted a comprehensive analysis of visual grounding of medical MLLMs.

Visual Grounding Analysis in General Domain MLLMs. Some recent studies have investigated the internal attention mechanisms of general-domain MLLMs, revealing their potential for implicit visual grounding. Zhang et al. (2025a) demonstrated that MLLMs can identify the correct spatial regions relevant to a given query, even without explicit grounding supervision. They introduce a training-free intervention method (e.g., cropping guided by attention or gradient maps) that enhances performance on general-domain VQA tasks. Broader research into MLLM interpretability has studied how visual information is fused into language representations. Techniques such as causal intervention and cross-modal attention visualization have been employed to offer insights into how vision and language tokens interact through attention mechanisms (Golovanevsky et al., 2024; Zhang et al., 2024; Yu & Ananiadou, 2025; Palit et al., 2023). These studies suggest that middle layers are especially crucial for integrating object-level visual cues with textual context, and that cross-modal attention patterns can encode meaningful spatial alignment signals. However, all of these insights have been drawn from general-domain visual data, such as natural scene images and standard VQA benchmarks. *In contrast, to our knowledge, no prior work has performed visual grounding analysis of medical MLLMs.*

B VGMED Scale Comparison with Related Attention Analysis Works

VGMED comprises approximately 28K image-bbox-question triplets, including 14K samples for localization questions and another 14K for attribute questions. The scale of VGMED is larger than or comparable to the number of samples used in the closely related RGB-domain visual grounding (Zhang et al., 2025a; Kang et al., 2025; Kaduri et al., 2024) / attention analysis work (Yang et al., 2025; Jiang et al., 2025; Chen et al., 2025a) (see Table B.1). Unlike RGB datasets that can be constructed by non-experts, our medical datasets require clinical expertise.

Table B.1: Number of samples used in related works.

Related works	**No. of Samples**	**Data Source**
MLLMs Know (Zhang et al., 2025a)	4,370	Text-VQA
Your LVLM (Kang et al., 2025)	1,000	RefCOCO
What's in the Image (Kaduri et al., 2024)	81	COCO
Hallucination Attribution (Yang et al., 2025)	1,500	COCO
Devils in LVLM (Jiang et al., 2025)	2,000	COCO
FastV (Chen et al., 2025a)	1,000	4 VL Tasks

C Detailed Information of Datasets Used in VGMED

Table C.2: Detailed information about the 44 datasets incorporated into VGMED. In the "Dataset" column, names such as "StructSeg2019 (Task 1)" represent specific task-based subsets. In the "Anatomical Structures" column, "Others" signifies datasets lacking detailed anatomical data from their original sources.

Dataset	**Modality**	**Anatomical Structures**
AMOS2022 (Ji et al., 2022)	CT, MR	Abdomen, Thorax, Pelvic
ATM2022 (Zhang et al., 2023a)	CT	Thorax
AbdomenomenCT-1K (Ma et al., 2022b)	CT	Abdomen
BTCV (Igelsias et al., 2015)	CT	Thorax, Abdomen, Pelvic
BraTS2013 (Menze et al., 2015)	MR	Head & neck
BraTS2015 (Menze et al., 2015)	MR	Head & neck
BraTS2018 (Menze et al., 2015)	MR	Head & neck
BraTS2019 (Menze et al., 2015)	MR	Head & neck
BraTS2020 (Menze et al., 2015)	MR	Head & neck
BraTS2021 (Bakas et al., 2017; Baid et al., 2021)	MR	Head & neck
CHAOS (Task 4) (Kavur et al., 2021)	MR	Abdomen
CTPelvic1k (Liu et al., 2021b)	CT	Pelvic
CVC-ClinicDB (Bernal et al., 2015)	Endoscopy	Others
Chest_Image_Pneum (Tianchi, 2020)	X-ray	Thorax
FLARE21 (Ma et al., 2022a)	CT	Abdomen
FLARE22 (Ma et al., 2024)	CT	Abdomen, Thorax
HVSMR2016 (Pace et al., 2015)	MR	Thorax
ADAM (Task 2) (Fang et al., 2022)	Fundus	Head & neck
PALM19 (Fu et al., 2019)	Fundus	Head & neck
ISLES (Maier et al., 2017)	MR	Head & neck
KiTS2019 (Heller et al., 2020)	CT	Abdomen
KiTS2021 (Zhao et al., 2021)	CT	Abdomen
LUNA16 (Setio et al., 2017)	CT	Thorax
MSD-BrainTumor (Antonelli et al., 2022)	MR	Head & neck
MSD-Liver (Antonelli et al., 2022)	CT	Abdomen
MSD-Pancreas (Antonelli et al., 2022)	CT	Abdomen
MSD-Spleen (Antonelli et al., 2022)	CT	Abdomen
CT-ORG (Antonelli et al., 2022)	CT	Head & neck, Thorax, Abdomen
PROMISE09 (Bharatha et al., 2001)	MR	Pelvic
PROMISE12 (Litjens et al., 2014)	MR	Pelvic
SIIM-ACR Pneumothorax (Zawacki et al., 2019)	X-ray	Thorax
StructSeg2019 (Task 1) (Huang et al., 2019)	CT	Head & neck
StructSeg2019 (Task 2) (Huang et al., 2019)	CT	Thorax, Abdomen
TotalSegmentator (Wasserthal et al., 2023)	CT	Head & neck, Thorax, Abdomen, Pelvic
Ultrasound Nerve Segmentation (Montoya et al., 2016)	Ultrasound	Others
WORD (Luo et al., 2022)	CT	Thorax, Abdomen
autoPET (Gatidis et al., 2022)	PET	Pelvic
BUSI (Al-Dhabyani et al., 2020)	Ultrasound	Thorax
Kvasir-SEG (Jha et al., 2019)	Endoscopy	Others
ISIC18 (Task 1) (Codella et al., 2019)	Dermoscopy	Skin
ISIC17 (Task 1) (Codella et al., 2018b)	Dermoscopy	Skin
ISIC16 (Task 1) (Codella et al., 2018a)	Dermoscopy	Skin
SLAKE (Liu et al., 2021a)	CT, MR, X-ray	Head & neck, Abdomen, Thorax
PolypDB (Jha et al., 2025)	Endoscopy	Others

D MORE DETAILS OF VGREFINE

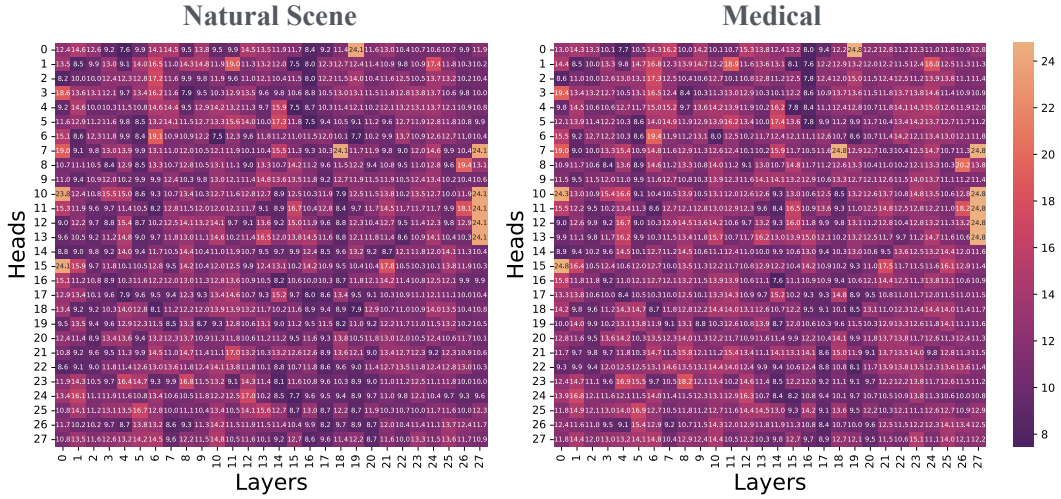

Figure D.1: We conduct an experiment to analyze the alignment between attention distributions from different attention heads and layers and the ground truth annotations in the images. This follows the evaluation setup described in Section 2.3 of the main paper. The medical MLLM evaluated is HuatuoGPT-V-7B. Each cell in the above figures reflects the degree of alignment, measured using our proposed KL Divergence metric (lower is better). *This analysis helps identify the specific heads and layers that are most relevant to visual grounding.* COCO is used for natural scene image analysis, and our dataset VGMED is used for medical image analysis. Interestingly, we find that the attention heads most relevant to visual grounding in natural scene images are often also the most relevant for medical images. However, despite this overlap, the overall visual grounding performance on medical images remains lower than on natural scenes, consistent with the findings presented in Figure 3 (main paper). Based on this analysis, we identify the top K attention heads with the strongest alignment (i.e., lowest KL divergence) and aggregate their attention distributions to compute a refined attention map. *Notably, we select the top K heads using randomly sampled natural scene images from COCO dataset, to avoid data leakage from medical evaluation benchmarks.* This setup also demonstrates that our method generalizes effectively from natural images to the biomedical domain.

In this section, we provide more details of our proposed inference-time method VGRefine (introduced in Sec. 3 of the main paper). Particularly, we discuss how we identify top K attention heads most relevant to visual grounding and leverage their attention distributions in Step I of VGRefine. Fig. D.1 depict the analysis.

We explore attention distributions from different attention heads across all layers, as prior work suggests that individual attention heads in transformers specialize in capturing distinct types of information Voita et al. (2019); Olsson et al. (2022); Gandelsman et al. (2024); Yu et al. (2023); Yang et al. (2025). This motivates us to examine attention at finer granularity to obtain the attention that focusing more on clinically relevant regions.

See details in Fig. D.1. Following the same evaluation setup of Sec. 2.3 of main paper, we assess relevancy to visual grounding of each attention head in HuatuoGPT-V by measuring the alignment between their attention distributions and ground-truth annotations. We perform this analysis using both natural scene images (from MS COCO) and medical images (from our VGMED). The alignment is measured by our proposed KL Divergence (↓) as metric.

As shown in Fig. D.1, the visual relevancy patterns are consistent across domains: heads that are relevant to visual grounding in natural scenes also show relative relevancy in medical images, despite exhibiting inadequate visual grounding on medical images compared to natural images (as discussed in Sec. 2.4). Based on this analysis, we select the top K heads with the highest visual grounding relevancy (lowest KL) on natural images and average their attention maps to obtain a refined attention map. This map is used in Step II to guide the model's improved focus on clinically meaningful areas.

E Experiments on Open-ended Medical VQA

We present additional experimental results on the open-ended questions from the Medical VQA benchmarks. Specifically, we evaluate on VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021a), and PathVQA (He et al., 2020), which include open-ended formats. As shown in Table E.3, our inference-time method consistently achieve better performance across all datasets, demonstrating its effectiveness in enhancing open-ended medical VQA.

Table E.3: Performance comparison on full medical VQA datasets for open-ended medical VQA. We evaluate all models under the zero-shot setting. These results underscore that enhanced visual grounding with our inference-time method VGRefine contributes to better performance on medical VQA tasks. It is important to note that VILA-M3 (Nath et al., 2024), MedPLIB (Huang et al., 2025) and LLaVA-Tri (Xie et al., 2025) incorporate training data from VQA-RAD, SLAKE, and PathVQA, and thus making zero-shot evaluation unfair, and are excluded from our zero-shot comparison.

Model	**VQA-RAD**				**SLAKE**				**PathVQA**			
Metric	BLEU-1	BERT	OpenRecall	Avg.	BLEU-1	BERT	OpenRecall	Avg.	BLEU-1	BERT	OpenRecall	Avg.
Qwen-VL-Chat	28.6	63.4	27.0	39.7	28.9	52.0	33.6	38.2	18.7	45.1	9.9	24.6
LLaVA-v1.6-7B	22.1	58.0	21.9	34.0	30.8	52.7	36.4	40.0	22.8	47.7	11.2	27.2
Med-Flamingo	27.4	61.9	12.7	34.0	11.8	40.2	21.1	24.4	24.3	50.4	2.4	25.7
RadFM	30.5	64.1	41.6	45.4	38.6	61.0	44.2	47.9	24.8	51.4	10.1	29.8
LLaVA-Med-7B	21.6	40.5	28.2	30.1	37.0	58.4	39.2	44.9	28.5	60.1	12.3	33.6
HuatuoGPT-V-7B	49.7	75.0	50.7	58.5	55.0	78.9	55.6	63.2	34.2	65.8	**36.5**	45.5
VGRefine-7B (Ours)	**51.2**	**76.3**	**52.3**	**59.9**	**56.5**	**80.0**	**56.7**	**64.4**	**36.1**	**68.1**	**36.5**	**46.9**

F Comparison with Other Attention-based Methods

We conducted an additional experiment comparing VGRefine with three very recent attention-based methods for medical MLLMs. Specifically, PAI (Liu et al., 2024c) and AdaptVis (Chen et al., 2025b) aim to refine/manipulate attention maps over visual tokens, while ViCrop (Zhang et al., 2025a) uses attention maps to enhance visual perception.

For a fair comparison, we implemented all methods on HuatuoGPT-V-7B, following their official code and hyperparameter settings. The experimental results, shown in Tab. F.4, indicate that VGRefine consistently outperforms all other methods.

Table F.4: Accuracy on closed-ended medical VQA datasets.

Model	**VQA-RAD**	**SLAKE**	**PathVQA**	**PMC-VQA**	**Avg.**
HuatuoGPT-V-7B (Baseline)	67.4	76.5	60.7	53.9	65.3
PAI(Liu et al., 2024c)	43.7	24.48	20.8	52.8	33.3
AdaptVis(Chen et al., 2025b)	68.6	75.1	**67.6**	52.9	66.7
ViCrop (Zhang et al., 2025a)	68.9	70.9	66.7	54.6	65.5
VGRefine (Ours)	**71.2**	**76.9**	**67.6**	**56.2**	**68.4**

G More Experiments on Larger Models

In this section, we provide more experimental results on larger models (with parameters $>$ 10B). We show comparison on all six benchmarks that are designed for biomedical MLLM evaluation, including VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021a), PathVQA (He et al., 2020), PMC-VQA (Zhang et al., 2023b), OmniMedVQA (Hu et al., 2024) (open-access split), and MMMU (Health & Medicine track) (Yue et al., 2024). All evaluations were conducted in a zero-shot setting using question templates provided by LLaVA (see Sec. I).

All experiments are conducted using the same hyperparameters across benchmarks. Specifically, for Step I, we aggregate the attention maps from the top $K = 20$ heads with the highest alignment to visual relevant regions, as measured by KL divergence on our curated evaluation set built using COCO images. This setup prevents data leakage from medical evaluation benchmarks and demonstrates that our method generalizes from natural images to biomedical domains. Low-activation regions

are suppressed based on a percentile threshold $p = 50\%$ over attention magnitude. For Step II we apply the attention knockout only at $\ell = 34, 35, 36$ layer, which, according to our analysis in Fig. 3 demonstrates the most relevancy to visual grounding among all the layers. We applied our inference-time method VGRefine-34B on HuatuoGPT-V-34B (Chen et al., 2024a). The hyperparameters K and p are kept consistent with the VGRefine-7B setting. In Step II, we apply attention knockout to more layers, as the 34B model has twice as many layers as the 7B variant and requires deeper intervention to achieve significant improvements.

Results in Tab. G.5 demonstrate that our proposed method consistently achieves good performance across all 6 benchmarks, demonstrating its effectiveness in enhancing all types of medical VQA.

Table G.5: Experiment results of larger models (more than 10B parameters). We evaluate all models under the zero-shot setting. Our inference-time method VGRefine outperforms other state-of-the-art medical MLLMs in most cases. These results underscore that enhanced visual grounding contributes to better performance on medical VQA tasks. It is important to note that VILA-M3 (Nath et al., 2024), MedRegA (Wang et al., 2025) incorporate training data from VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021a), PathVQA (He et al., 2020), and PMC-VQA (Zhang et al., 2023b), thus making zero-shot evaluation unfair, and are excluded from our zero-shot comparison of these benchmarks.

Benchmarks	**Subset**	**Metric**	**LLaVA-v1.6-34B**	**VILA-M3-13B**	**MedRegA-34B**	**HuatuoGPT-V-34B**	**VGRefine-34B** (Ours)
VQA-RAD	-	CloseAcc	58.6	-	-	68.1	**72.9**
		BLEU-1	44.5	-	-	50.5	**52.6**
		BERT	69.2	-	-	**74.8**	**74.8**
		OpenRecall	43.6	-	-	51.7	**52.8**
SLAKE	-	CloseAcc	67.3	-	-	76.9	**79.1**
		BLEU-1	48.6	-	-	56.3	**57.2**
		BERT	51.8	-	-	77.6	**79.5**
		OpenRecall	54.2	-	-	57.5	**58.5**
PathVQA	-	CloseAcc	59.1	-	-	63.5	**69.7**
		BLEU-1	28.1	-	-	36.6	**37.6**
		BERT	57.7	-	-	65.6	**65.9**
		OpenRecall	29.3	-	-	36.9	**37.1**
PathVQA	-	CloseAcc	44.4	-	-	58.2	**58.7**
Avg. on Med-VQAs	-	CloseAcc	57.4	-	-	67.0	**70.7**
MMMU	BMS	CloseAcc	56.4	36.8	54.3	64.3	**66.0**
	CM		52.8	38.8	53.5	56.5	**58.2**
	DLM		42.6	29.0	37.7	45.1	**45.4**
	P		41.6	29.3	38.4	43.7	**44.0**
	PH		38.4	32.2	40.7	43.8	**44.8**
	Avg.		45.6	33.3	44.7	50.1	**51.3**
OmniMedVQA	CT	CloseAcc	50.6	56.9	62.5	69.7	**71.7**
	FP		63.4	50.1	80.4	**84.6**	84.4
	MRI		60.9	52.9	72.7	69.7	**73.9**
	OCT		68.4	41.5	86.2	**87.8**	87.6
	Der		65.7	45.1	**79.9**	70.2	70.9
	Mic		62.8	50.6	71.3	71.1	**71.4**
	X-Ray		74.7	62.5	78.7	83.8	**84.7**
	US		44.5	47.1	49.4	81.7	**83.1**
	Avg.		61.4	52.3	70.3	74.4	**76.6**

H HuatuoGPT-Vision-Bio with BiomedCLIP Vision Encoder

Model Setup. To evaluate the effect of domain-specific visual encoders, we modified the original HuatuoGPT-Vision architecture by replacing its CLIP-based vision encoder with BioMedCLIP Zhang et al. (2025b), a biomedical foundation model pretrained on 15 million scientific image–text pairs. All other components of the model (including the Qwen2 language model, the cross-modal connector module, and the training protocol) remain identical to the original configuration. *This substitution allows us to isolate the impact of specialized medical image representations on visual grounding performance.*

Training Details. Since the original training code for HuatuoGPT-Vision was not publicly available, we replicated the training pipeline using the LLaVA-NeXT Liu et al. (2024b) codebase. We follow a two-stage training protocol on the same pretraining and instruction-tuning datasets used in HuatuoGPT-Vision, including LLaVA and PubMedVision. In Stage I, we freeze both the BioMedCLIP vision encoder and the Qwen2 language model, training only the connector to align visual and textual representations. In Stage II, we fine-tune both the connector and the language model while keeping the vision encoder frozen. The model is trained for 1 epoch. BioMedCLIP processes images at a fixed resolution of 224×224 with a patch size of 16, which differs from the resolution and tokenization settings used in the original CLIP-based HuatuoGPT-Vision.

Analysis Results. As shown in main paper Fig. 1 and Fig. 3, *the issue of suboptimal visual grounding on medical images cannot be solved by using BiomedCLIP vision encoder.*

I PROMPTS FOR VGMED AND QA EVALUATION

I.1 PROMPTS FOR CONSTRUCTING VGMED

Localization Question Set

- `Is there a {label} in the image?`
- `Can you see a {label} in the image?`
- `Does the image contain a {label}?`
- `Is a {label} present in this image?`
- `Do you see a {label} in the picture?`
- `Is the {label} visible in the image?`
- `Is there any sign of a {label} in the image?`
- `Can a {label} be found in this image?`
- `Does this image show a {label}?`
- `Is a {label} shown in the picture?`

Figure I.2: The question from a predefined question set is sampled for generating localization questions in both the COCO and VGMED datasets. {label} represents the object (in COCO) or organ/lesion (in VGMED) identified by a bounding box in the corresponding image.

I.2 PROMPTS FOR ZERO-SHOT EVALUATION

We used the LLaVA prompt template during the evaluation for open, closed-ended, and multiple-choice questions.

Short Answer (e.g., VQA-RAD, SLAKE, PathVQA)

```
<question>
Answer the question using a single word or phrase.
```

Figure I.12: Prompt for evaluating the open and closed-ended questions in VQA-RAD, SLAKE, and PathVQA benchmarks.

Option-only for multiple-choice (e.g., PMC-VQA, OmniMedVQA, and MMMU)

```
<question>
   A. <option_1>
   B. <option_2>
   C. <option_3>
   D. <option_4>
Answer with the option's letter from the given choices
directly.
```

Figure I.13: Prompt for evaluating the multiple-choice VQA benchmarks.

Prompt for VGMED Attribute Questions (MRI)

Your task is to generate clinically meaningful questions for an evaluation dataset to assess visual grounding capability in medical reasoning, without requiring deep semantic grounding.

- Semantic Grounding: Anchors linguistic representations to domain-specific medical knowledge to inform what to look for, ensuring accurate reasoning about diseases or anatomical concepts.
- Visual Grounding: Localizes and interprets specific regions in medical images based on relevant features, enabling spatial alignment of language queries to visual elements for accurate analysis.

We will provide the label of a medical structure (organ, lesion, or tissue) in a {modality} image, derived from a bounding box annotation. Your need to generate three clinically relevant questions about visual attributes of this structure.

Guidelines for the question:

- Focus on visual grounding, without requiring deep medical semantic grounding.
- Ensure clinical relevance.
- Require attention to the entire annotated bounding box.
- Address only observable visual characteristics (e.g., size, shape, density, enhancement, homogeneity).
- Avoid referencing other body parts or surrounding structures.
- Do not include position, modality, or plane.
- Exclude diagnoses or treatments requiring deep semantic grounding.
- Avoid compound or multi-condition questions.
- Ensure variety across the three questions.

Example questions:

- "Is the lesion hyper or hypointense?"
- "Is the lesion enhancing?"
- "What does the area of necrosis look like?"
- "What pattern of enhancement does the lesion show?"

Now, generate three questions based on the label. Return exactly three questions without any additional text or formatting.

Label: {label}

Figure I.3: For attribute questions in VGMED, we use a specific prompt for each modality. In the prompt, {modality} denotes the modality of the image. {label} denotes the organ or lesion labeled by a bounding box in the image.

Prompt for VGMED Attribute Questions (CT)

Your task is to generate clinically meaningful questions for an evaluation dataset to assess visual grounding capability in medical reasoning, without requiring deep semantic grounding.

- Semantic Grounding: Anchors linguistic representations to domain-specific medical knowledge to inform what to look for, ensuring accurate reasoning about diseases or anatomical concepts.
- Visual Grounding: Localizes and interprets specific regions in medical images based on relevant features, enabling spatial alignment of language queries to visual elements for accurate analysis.

We will provide the label of a medical structure (organ, lesion, or tissue) in a {modality} image, derived from a bounding box annotation. Your need to generate three clinically relevant questions about visual attributes of this structure.

Guidelines for the question:

- Focus on visual grounding, without requiring deep medical semantic grounding.
- Ensure clinical relevance.
- Require attention to the entire annotated bounding box.
- Address only observable visual characteristics (e.g., size, shape, density, enhancement, homogeneity).
- Avoid referencing other body parts or surrounding structures.
- Do not include position, modality, or plane.
- Exclude diagnoses or treatments requiring deep semantic grounding.
- Avoid compound or multi-condition questions.
- Ensure variety across the three questions.

Example questions:

- "Are there ground glass opacities within the lung?"
- "Is the kidney enlarged?"
- "What is the size of the necrosis?"

Now, generate three questions based on the label. Return exactly three questions without any additional text or formatting.

Label: {label}

Figure I.4: For attribute questions in VGMED, we use a specific prompt for each modality. In the prompt, {modality} denotes the modality of the image. {label} denotes the organ or lesion labeled by a bounding box in the image.

Prompt for VGMED Attribute Questions (Ultrasound)

Your task is to generate clinically meaningful questions for an evaluation dataset to assess visual grounding capability in medical reasoning, without requiring deep semantic grounding.

- Semantic Grounding: Anchors linguistic representations to domain-specific medical knowledge to inform what to look for, ensuring accurate reasoning about diseases or anatomical concepts.
- Visual Grounding: Localizes and interprets specific regions in medical images based on relevant features, enabling spatial alignment of language queries to visual elements for accurate analysis.

We will provide the label of a medical structure (organ, lesion, or tissue) in a {modality} image, derived from a bounding box annotation. Your need to generate three clinically relevant questions about visual attributes of this structure.

Guidelines for the question:

- Focus on visual grounding, without requiring deep medical semantic grounding.
- Ensure clinical relevance.
- Require attention to the entire annotated bounding box.
- Address only observable visual characteristics (e.g., size, shape, density, enhancement, homogeneity).
- Avoid referencing other body parts or surrounding structures.
- Do not include position, modality, or plane.
- Exclude diagnoses or treatments requiring deep semantic grounding.
- Avoid compound or multi-condition questions.
- Ensure variety across the three questions.

Example questions:

- "Does the thyroid nodule have irregular or microlobulated margins?"
- "Does the thyroid nodule have marked hypoechogenicity?"
- "Does the thyroid nodule have multiple microcalcifications?"
- "Is the breast lesion homogeneous or heterogeneous?"
- "Does the breast lesion appear solid or cystic on ultrasound?"

Now, generate three questions based on the label. Return exactly three questions without any additional text or formatting.

Label: {label}

Figure I.5: For attribute questions in VGMED, we use a specific prompt for each modality. In the prompt, {modality} denotes the modality of the image. {label} denotes the organ or lesion labeled by a bounding box in the image.

Prompt for VGMED Attribute Questions (X-ray)

Your task is to generate clinically meaningful questions for an evaluation dataset to assess visual grounding capability in medical reasoning, without requiring deep semantic grounding.

- Semantic Grounding: Anchors linguistic representations to domain-specific medical knowledge to inform what to look for, ensuring accurate reasoning about diseases or anatomical concepts.
- Visual Grounding: Localizes and interprets specific regions in medical images based on relevant features, enabling spatial alignment of language queries to visual elements for accurate analysis.

We will provide the label of a medical structure (organ, lesion, or tissue) in a {modality} image, derived from a bounding box annotation. Your need to generate three clinically relevant questions about visual attributes of this structure.

Guidelines for the question:

- Focus on visual grounding, without requiring deep medical semantic grounding.
- Ensure clinical relevance.
- Require attention to the entire annotated bounding box.
- Address only observable visual characteristics (e.g., size, shape, density, enhancement, homogeneity).
- Avoid referencing other body parts or surrounding structures.
- Do not include position, modality, or plane.
- Exclude diagnoses or treatments requiring deep semantic grounding.
- Avoid compound or multi-condition questions.
- Ensure variety across the three questions.

Example questions:

- "What is the size of the pneumothorax?"
- "Where is the pneumothorax?"
- "Does the lung field appear more opaque or translucent in the annotated region?"

Now, generate three questions based on the label. Return exactly three questions without any additional text or formatting.

Label: {label}

Figure I.6: For attribute questions in VGMED, we use a specific prompt for each modality. In the prompt, {modality} denotes the modality of the image. {label} denotes the organ or lesion labeled by a bounding box in the image.

Prompt for VGMED Attribute Questions (Fundus Photography)

Your task is to generate clinically meaningful questions for an evaluation dataset to assess visual grounding capability in medical reasoning, without requiring deep semantic grounding.

- Semantic Grounding: Anchors linguistic representations to domain-specific medical knowledge to inform what to look for, ensuring accurate reasoning about diseases or anatomical concepts.
- Visual Grounding: Localizes and interprets specific regions in medical images based on relevant features, enabling spatial alignment of language queries to visual elements for accurate analysis.

We will provide the label of a medical structure (organ, lesion, or tissue) in a {modality} image, derived from a bounding box annotation. Your need to generate three clinically relevant questions about visual attributes of this structure.

Guidelines for the question:

- Focus on visual grounding, without requiring deep medical semantic grounding.
- Ensure clinical relevance.
- Require attention to the entire annotated bounding box.
- Address only observable visual characteristics (e.g., size, shape, density, enhancement, homogeneity).
- Avoid referencing other body parts or surrounding structures.
- Do not include position, modality, or plane.
- Exclude diagnoses or treatments requiring deep semantic grounding.
- Avoid compound or multi-condition questions.
- Ensure variety across the three questions.

Example questions:

- "Is there any pallor observed in the optic disc?"
- "Does the optic disc appear swollen or elevated?"
- "Is there any evidence of swelling or pallor in the optic disc?"

Now, generate three questions based on the label. Return exactly three questions without any additional text or formatting.

Label: {label}

Figure I.7: For attribute questions in VGMED, we use a specific prompt for each modality. In the prompt, {modality} denotes the modality of the image. {label} denotes the organ or lesion labeled by a bounding box in the image.

Prompt for VGMED Attribute Questions (Endoscopy)

Your task is to generate clinically meaningful questions for an evaluation dataset to assess visual grounding capability in medical reasoning, without requiring deep semantic grounding.

- Semantic Grounding: Anchors linguistic representations to domain-specific medical knowledge to inform what to look for, ensuring accurate reasoning about diseases or anatomical concepts.
- Visual Grounding: Localizes and interprets specific regions in medical images based on relevant features, enabling spatial alignment of language queries to visual elements for accurate analysis.

We will provide the label of a medical structure (organ, lesion, or tissue) in a {modality} image, derived from a bounding box annotation. Your need to generate three clinically relevant questions about visual attributes of this structure.

Guidelines for the question:

- Focus on visual grounding, without requiring deep medical semantic grounding.
- Ensure clinical relevance.
- Require attention to the entire annotated bounding box.
- Address only observable visual characteristics (e.g., size, shape, density, enhancement, homogeneity).
- Avoid referencing other body parts or surrounding structures.
- Do not include position, modality, or plane.
- Exclude diagnoses or treatments requiring deep semantic grounding.
- Avoid compound or multi-condition questions.
- Ensure variety across the three questions.

Example questions:

- "What is the size of the polyp?"
- "Does the colorectal polyp have a smooth or lobulated surface appearance?"
- "What is the mobility of the polyp?"

Now, generate three questions based on the label. Return exactly three questions without any additional text or formatting.

Label: {label}

Figure I.8: For attribute questions in VGMED, we use a specific prompt for each modality. In the prompt, {modality} denotes the modality of the image. {label} denotes the organ or lesion labeled by a bounding box in the image.

Prompt for VGMED Attribute Questions (PET)

Your task is to generate clinically meaningful questions for an evaluation dataset to assess visual grounding capability in medical reasoning, without requiring deep semantic grounding.

- Semantic Grounding: Anchors linguistic representations to domain-specific medical knowledge to inform what to look for, ensuring accurate reasoning about diseases or anatomical concepts.
- Visual Grounding: Localizes and interprets specific regions in medical images based on relevant features, enabling spatial alignment of language queries to visual elements for accurate analysis.

We will provide the label of a medical structure (organ, lesion, or tissue) in a {modality} image, derived from a bounding box annotation. Your need to generate three clinically relevant questions about visual attributes of this structure.

Guidelines for the question:

- Focus on visual grounding, without requiring deep medical semantic grounding.
- Ensure clinical relevance.
- Require attention to the entire annotated bounding box.
- Address only observable visual characteristics (e.g., size, shape, density, enhancement, homogeneity).
- Avoid referencing other body parts or surrounding structures.
- Do not include position, modality, or plane.
- Exclude diagnoses or treatments requiring deep semantic grounding.
- Avoid compound or multi-condition questions.
- Ensure variety across the three questions.

Example questions:

- "Does the lesion show increased radiotracer uptake on the PET scan?"
- "Is the lesion hypo- or hyper-metabolic?"
- "What is the Standardized Uptake Value (SUV) of the lesion?"

Now, generate three questions based on the label. Return exactly three questions without any additional text or formatting.

Label: {label}

Figure I.9: For attribute questions in VGMED, we use a specific prompt for each modality. In the prompt, {modality} denotes the modality of the image. {label} denotes the organ or lesion labeled by a bounding box in the image.

Prompt for VGMED Attribute Questions (Dermoscopy)

Your task is to generate clinically meaningful questions for an evaluation dataset to assess visual grounding capability in medical reasoning, without requiring deep semantic grounding.

- Semantic Grounding: Anchors linguistic representations to domain-specific medical knowledge to inform what to look for, ensuring accurate reasoning about diseases or anatomical concepts.
- Visual Grounding: Localizes and interprets specific regions in medical images based on relevant features, enabling spatial alignment of language queries to visual elements for accurate analysis.

We will provide the label of a medical structure (organ, lesion, or tissue) in a {modality} image, derived from a bounding box annotation. Your need to generate three clinically relevant questions about visual attributes of this structure.

Guidelines for the question:

- Focus on visual grounding, without requiring deep medical semantic grounding.
- Ensure clinical relevance.
- Require attention to the entire annotated bounding box.
- Address only observable visual characteristics (e.g., size, shape, density, enhancement, homogeneity).
- Avoid referencing other body parts or surrounding structures.
- Do not include position, modality, or plane.
- Exclude diagnoses or treatments requiring deep semantic grounding.
- Avoid compound or multi-condition questions.
- Ensure variety across the three questions.

Example questions:

- "What is the size of the lesion?"
- "Is the lesion hypo- or hyper pigmented?"
- "Does the lesion have peripheral black dots or clods?"
- "Does the skin lesion have thick lines (reticular or branched)?"
- "Does the lesion have Polymorphous vessels?"

Now, generate three questions based on the label. Return exactly three questions without any additional text or formatting.

Label: {label}

Figure I.10: For attribute questions in VGMED, we use a specific prompt for each modality. In the prompt, {modality} denotes the modality of the image. {label} denotes the organ or lesion labeled by a bounding box in the image.

Prompt for COCO Attribute Questions

Your task is to generate one simple and meaningful question about a visual attribute of an object identified in an image.

We will provide the label of the object, which comes from a bounding box annotation in an image from the COCO dataset.

Guidelines for the question:

- Ensure variety in the questions generated.
- Focus only on the visual characteristics (e.g., color, size, material, etc.) of the given object.
- Do not reference other parts of the image.
- Do not ask questions about the position of the object or the surrounding structure.
- Avoid compound or multi-condition questions.

Now, generate one question based on the following label:
Label: {label}

Figure I.11: In order to compare the results between our VGMED datasets and natural scene images, we have also generated the attribute questions for COCO examples. {label} refers to the object label from the image's bounding boxes.

J Additional Qualitative Evaluation

J.1 Human Evaluation

We conducted a blinded human evaluation involving five experienced clinicians (4 of them have over 10 years of clinical practice). The study was based on a 20-case questionnaire. For each case, clinicians were shown a medical image with a VQA question and two corresponding attention maps: (1) from the baseline model and (2) from the same model after applying VGRefine. The source of each attention map was not disclosed, and their order was randomized. Clinicians were asked: "*Which model's attention visualization (shown as heatmap) appears more clinically reasonable and trustworthy?*".

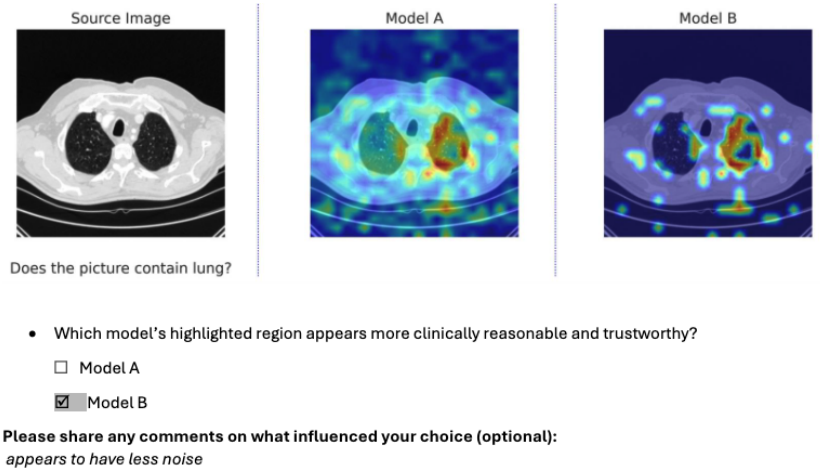

Figure J.14: Example of a blinded human evaluation case, showing a medical image with a VQA question, baseline attention map, and VGRefine attention map, assessed by an experienced clinician for clinical reasonableness. Clinician feedback highlighted that VGRefine attention maps were less noisy, better localized, and more aligned with expected clinical focus points.

J.2 Additional Qualitative Analysis on Medical MLLM's Attention Maps

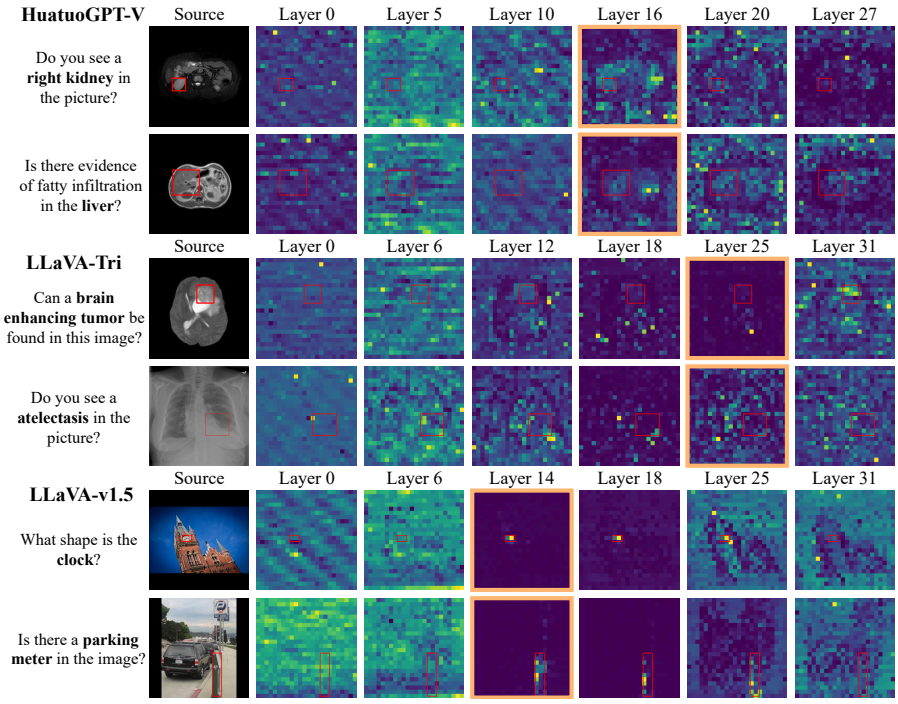

Figure J.15: **Qualitative evaluation.** We visualize attention maps across different layers, including those with the lowest KL divergence (highlighted with an orange boundary), which are indicative of layers most relevant to visual grounding in MLLMs. For medical images, the attention maps of medical MLLMs show limited alignment with the ground-truth annotated regions. In contrast, a general-domain MLLM LLaVA-v1.5 applied to natural images exhibits strong alignment with relevant regions, consistent with other study of general-domain MLLMs (Zhang et al., 2025a). This highlights a gap in MLLM's visual grounding performance between the medical and natural image domains. Best viewed in color and with zoom. **Additional results in Supp J.2**.

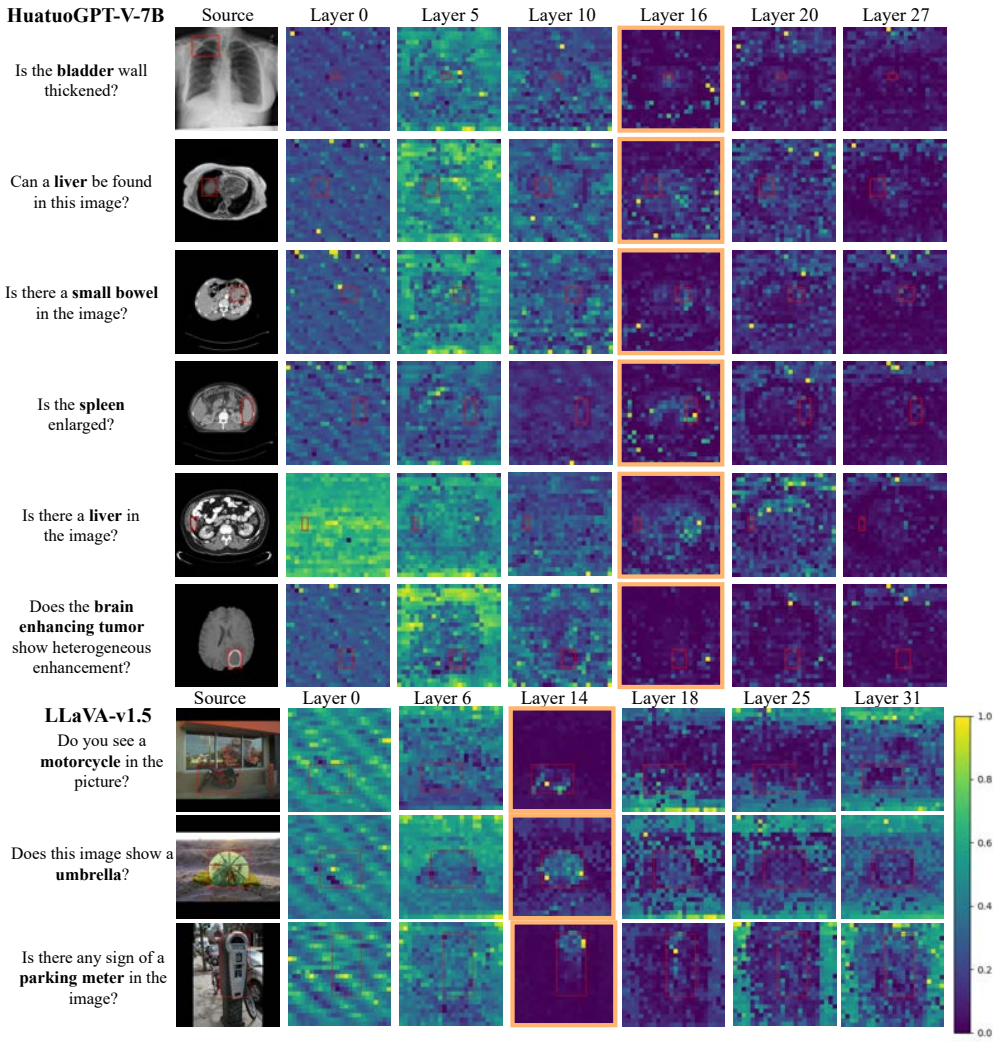

Figure J.16: **Qualitative evaluation.** We visualize attention maps across different layers, including those with the lowest KL divergence (highlighted with an orange boundary), which are indicative of layers most relevant to visual grounding in MLLMs. For medical images, the attention maps of medical MLLMs show limited alignment with the ground-truth annotated regions. In contrast, a general-domain MLLM LLaVA-v1.5 applied to natural images exhibits strong alignment with relevant regions, consistent with other study of general-domain MLLMs Zhang et al. (2025a). This highlights a gap in MLLM's visual grounding performance between the medical and natural image domains.

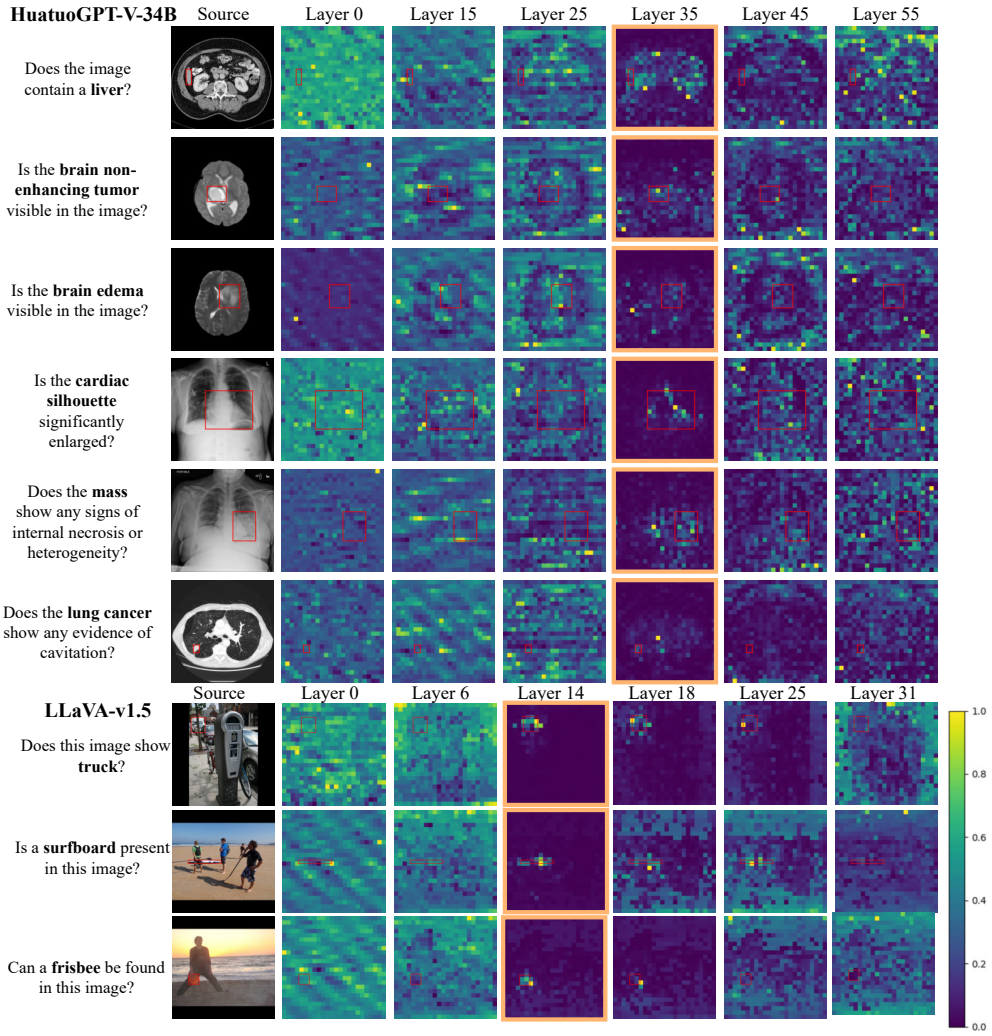

Figure J.17: **Qualitative evaluation.** We visualize attention maps across different layers, including those with the lowest KL divergence (highlighted with an orange boundary), which are indicative of layers most relevant to visual grounding in MLLMs. For medical images, the attention maps of medical MLLMs show limited alignment with the ground-truth annotated regions. In contrast, a general-domain MLLM LLaVA-v1.5 applied to natural images exhibits strong alignment with relevant regions, consistent with other study of general-domain MLLMs Zhang et al. (2025a). This highlights a gap in MLLM's visual grounding performance between the medical and natural image domains.

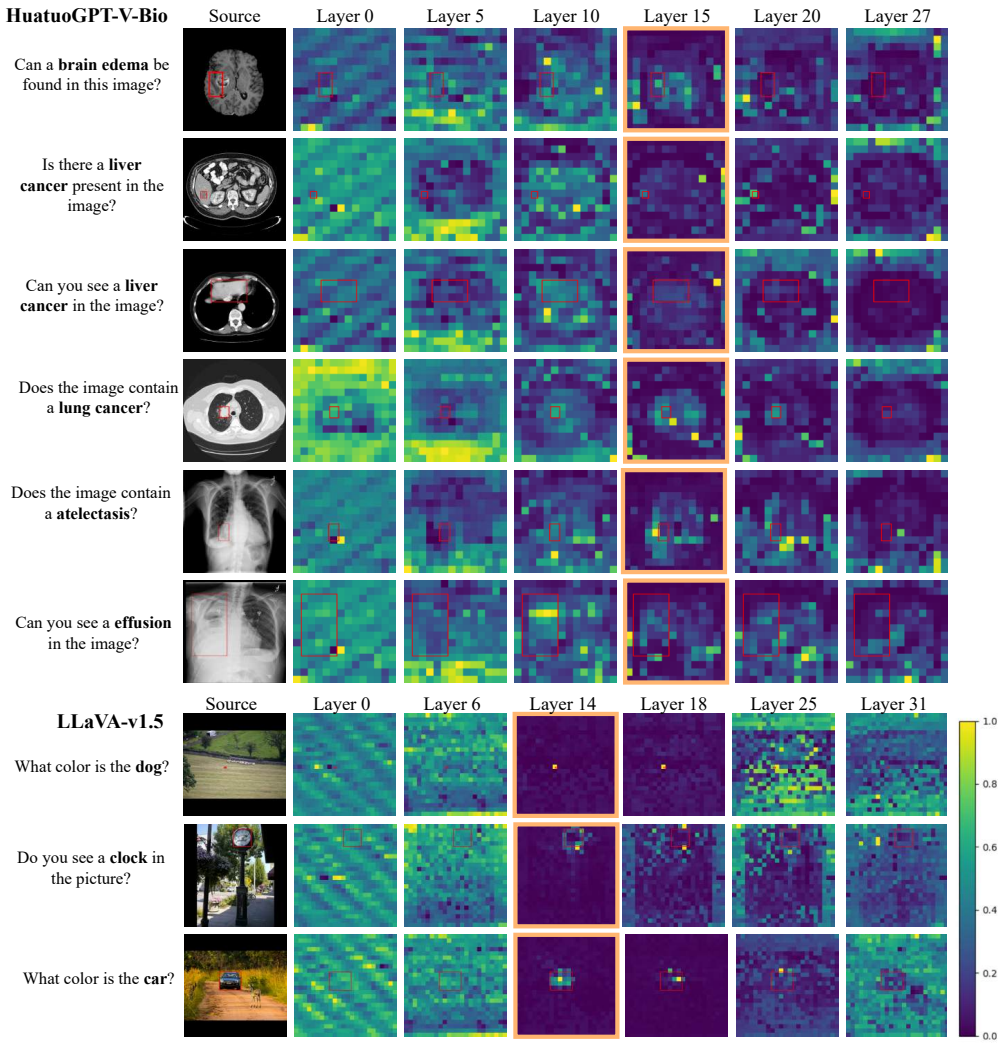

Figure J.18: **Qualitative evaluation.** We visualize attention maps across different layers, including those with the lowest KL divergence (highlighted with an orange boundary), which are indicative of layers most relevant to visual grounding in MLLMs. For medical images, the attention maps of HuatuoGPT-V-Bio with a specialized vision encoder (BiomedCLIP) show limited alignment with the ground-truth annotated regions. In contrast, a general-domain MLLM LLaVA-v1.5 applied to natural images exhibits strong alignment with relevant regions, consistent with other study of general-domain MLLMs Zhang et al. (2025a).

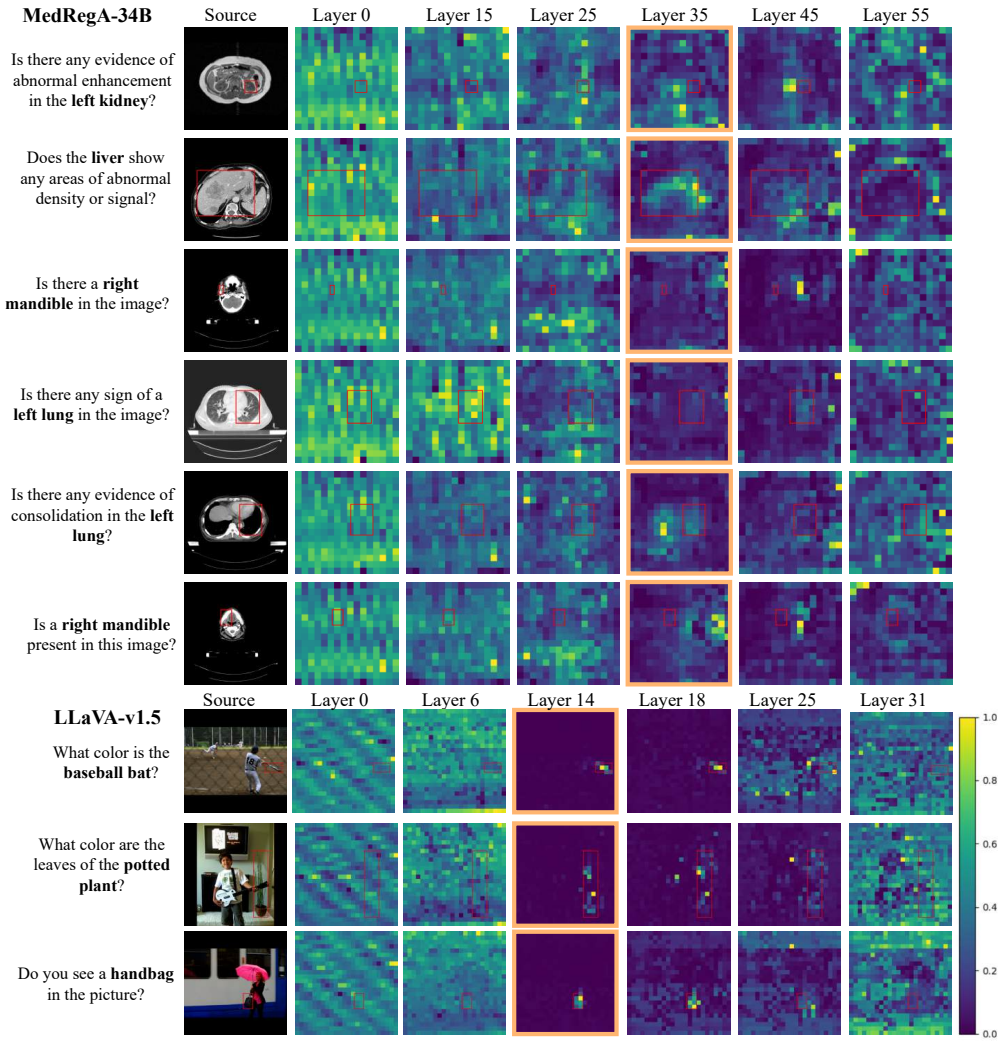

Figure J.19: **Qualitative evaluation.** We visualize attention maps across different layers, including those with the lowest KL divergence (highlighted with an orange boundary), which are indicative of layers most relevant to visual grounding in MLLMs. For medical images, the attention maps of medical MLLMs show limited alignment with the ground-truth annotated regions. In contrast, a general-domain MLLM LLaVA-v1.5 applied to natural images exhibits strong alignment with relevant regions, consistent with other study of general-domain MLLMs Zhang et al. (2025a). This highlights a gap in MLLM's visual grounding performance between the medical and natural image domains.

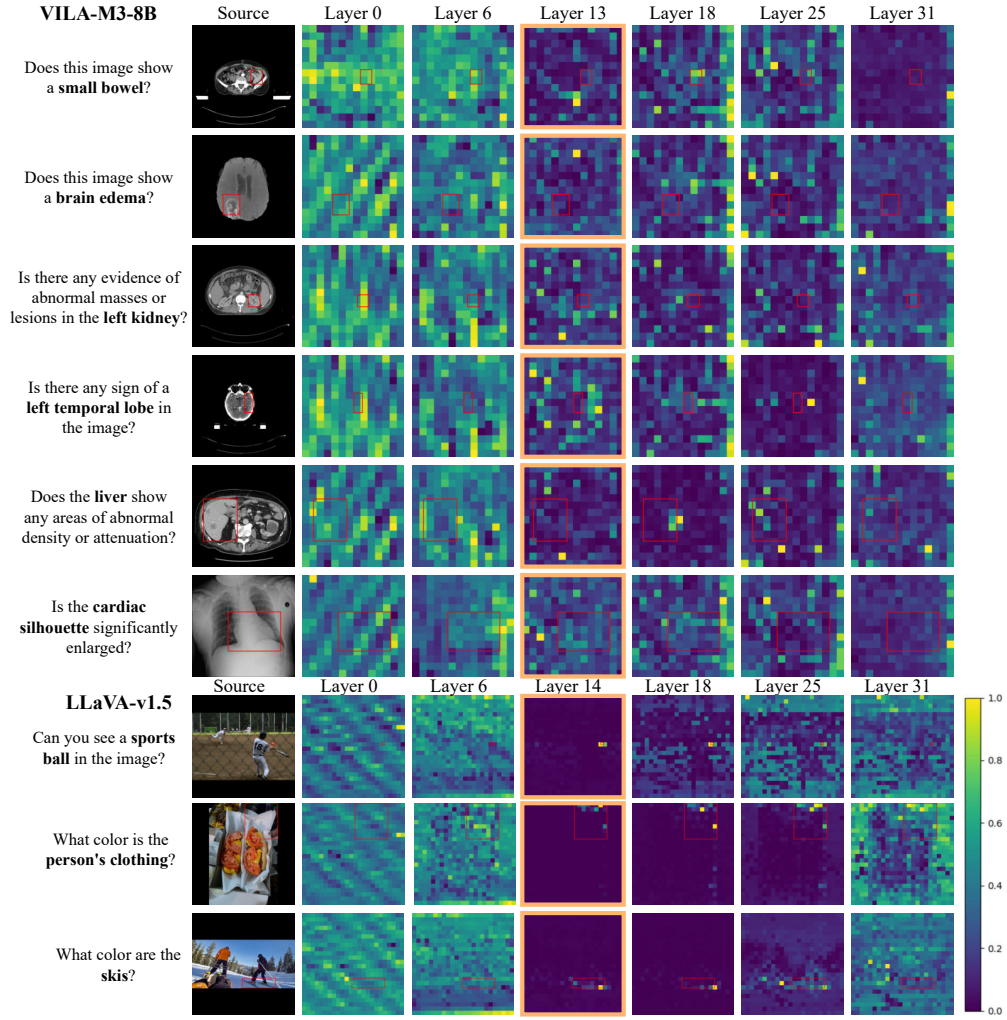

Figure J.20: **Qualitative evaluation.** We visualize attention maps across different layers, including those with the lowest KL divergence (highlighted with an orange boundary), which are indicative of layers most relevant to visual grounding in MLLMs. For medical images, the attention maps of medical MLLMs show limited alignment with the ground-truth annotated regions. In contrast, a general-domain MLLM LLaVA-v1.5 applied to natural images exhibits strong alignment with relevant regions, consistent with other study of general-domain MLLMs Zhang et al. (2025a). This highlights a gap in MLLM's visual grounding performance between the medical and natural image domains.

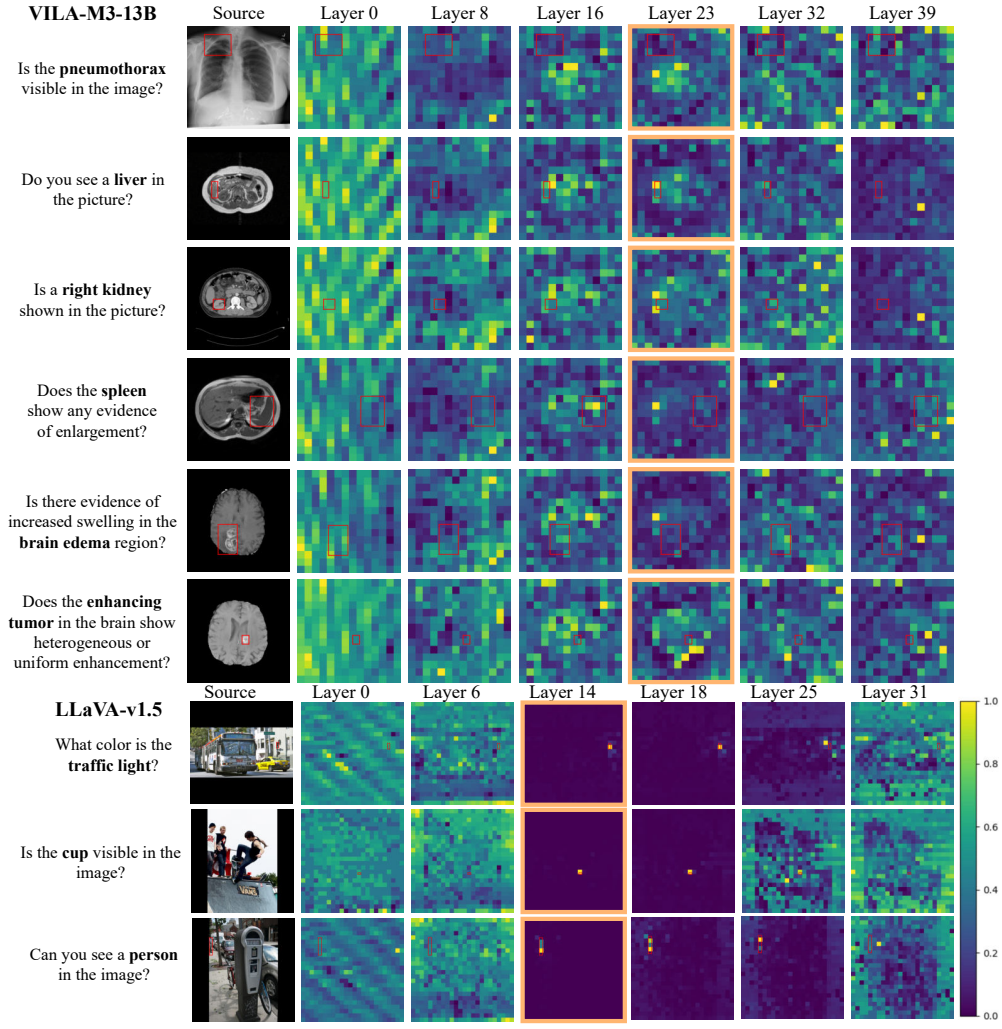

Figure J.21: **Qualitative evaluation.** We visualize attention maps across different layers, including those with the lowest KL divergence (highlighted with an orange boundary), which are indicative of layers most relevant to visual grounding in MLLMs. For medical images, the attention maps of medical MLLMs show limited alignment with the ground-truth annotated regions. In contrast, a general-domain MLLM LLaVA-v1.5 applied to natural images exhibits strong alignment with relevant regions, consistent with other study of general-domain MLLMs Zhang et al. (2025a). This highlights a gap in MLLM's visual grounding performance between the medical and natural image domains.

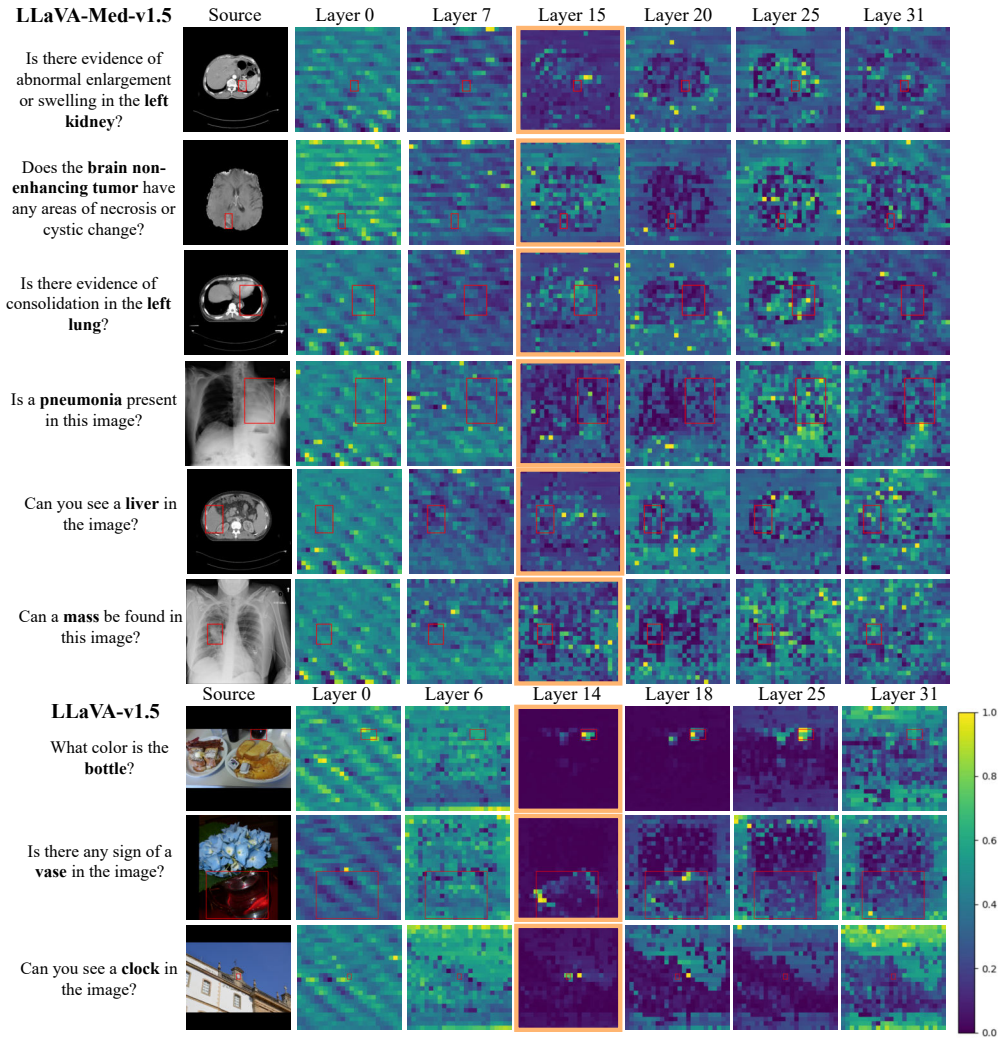

Figure J.22: **Qualitative evaluation.** We visualize attention maps across different layers, including those with the lowest KL divergence (highlighted with an orange boundary), which are indicative of layers most relevant to visual grounding in MLLMs. For medical images, the attention maps of medical MLLMs show limited alignment with the ground-truth annotated regions. In contrast, a general-domain MLLM LLaVA-v1.5 applied to natural images exhibits strong alignment with relevant regions, consistent with other study of general-domain MLLMs Zhang et al. (2025a). This highlights a gap in MLLM's visual grounding performance between the medical and natural image domains.

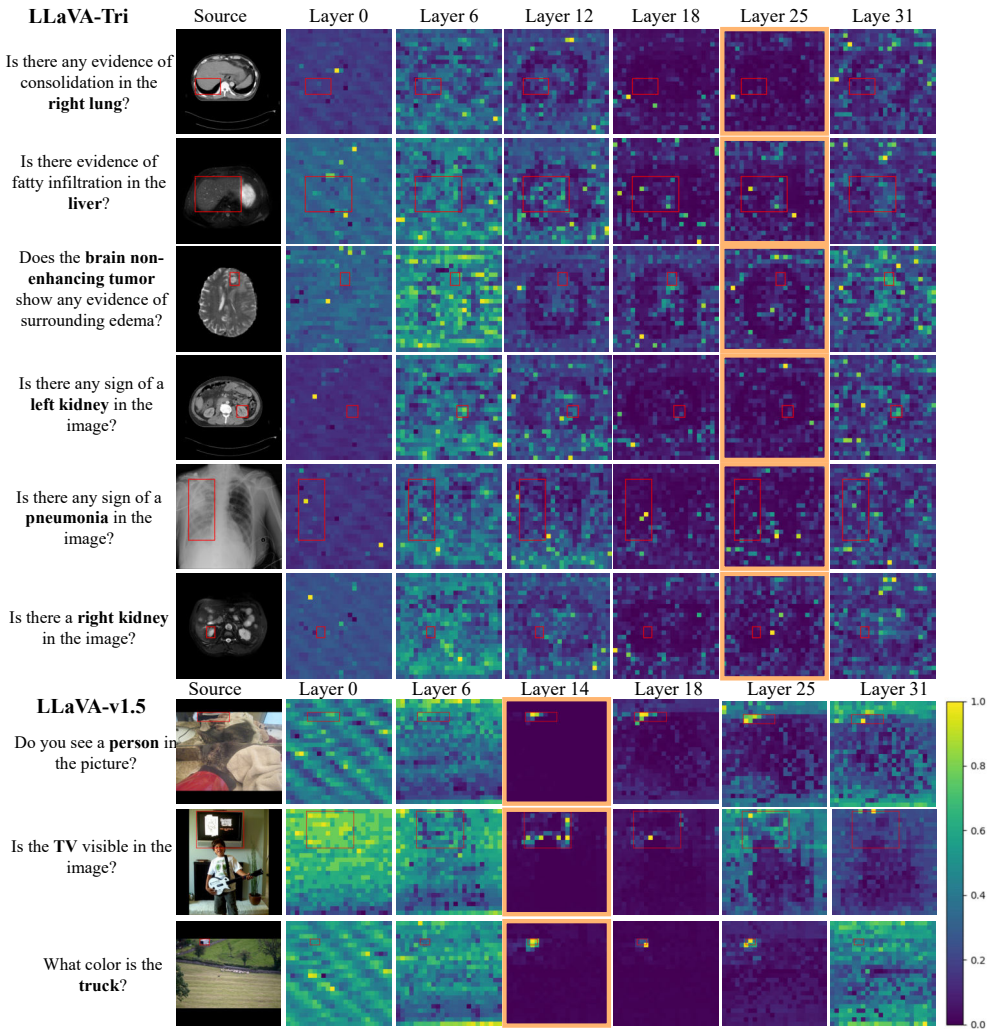

Figure J.23: **Qualitative evaluation.** We visualize attention maps across different layers, including those with the lowest KL divergence (highlighted with an orange boundary), which are indicative of layers most relevant to visual grounding in MLLMs. For medical images, the attention maps of medical MLLMs show limited alignment with the ground-truth annotated regions. In contrast, a general-domain MLLM LLaVA-v1.5 applied to natural images exhibits strong alignment with relevant regions, consistent with other study of general-domain MLLMs Zhang et al. (2025a). This highlights a gap in MLLM's visual grounding performance between the medical and natural image domains.

K LIMITATIONS

While our work provides a systematic and detailed investigation into visual grounding as a key failure mode in medical MLLMs, it focuses exclusively on this aspect. We do not examine other potential sources of failure, such as deficiencies in semantic grounding or reasoning capabilities. In practice, failures may also arise from an inability to recognize what clinical concepts are relevant or to integrate multimodal information effectively. Additionally, our proposed method, VGRefine, is designed to improve visual grounding at inference time but does not address other underlying limitations, such as dataset biases or insufficient domain-specific knowledge. Future work will explore complementary methods to assess and improve semantic grounding and extend our analysis framework to uncover other failure modes.

L EXPERIMENTAL SETTING/DETAILS AND COMPUTING RESOURCES

For both VGRefine-7B and VGRefine-34B, we select the top 20 attention heads—ranked by alignment with visually relevant regions—and average their outputs to obtain the filtered attention map. A percentile threshold of 50% is used to suppress low-activation regions during attention knockout. For VGRefine-7B, the attention knockout is applied at layer $l = 16$, while for VGRefine-34B, it is applied at layers $l = 34, 35, 36$, identified through our quantitative analysis as most relevant to visual grounding. We follow a zero-shot evaluation protocol across six biomedical VQA benchmarks: VQA-RAD, SLAKE, PathVQA, PMC-VQA, OmniMedVQA, and MMMU (Health & Medicine track). The full set of prompts used for zero-shot evaluation is provided in Section F.2. All experiments are conducted on a server with 8×NVIDIA A100 80GB GPUs.

M BROADER IMPACTS AND ETHICAL CONSIDERATIONS

This work involves the analysis of medical MLLMs using publicly available datasets that are de-identified. No private or sensitive patient data is used. We acknowledge that the deployment of medical MLLMs carries potential risks, including misinterpretation of clinical images, over-reliance on automated outputs by clinicians, and disparities in performance across patient populations. Our work aims to mitigate such risks by improving the reliability of model predictions through better visual grounding. To promote transparency and reproducibility, we provide open access to code, evaluation metrics, and the VGMED dataset. This enables the broader research community to scrutinize and build upon our work responsibly.

N SAFEGUARDS

Our study does not involve training or releasing a new foundation model, but rather evaluates and analyzes existing medical MLLMs in terms of their visual grounding behavior. While our proposed inference-time refinement method improves grounding performance, it is designed for research use only and does not replace expert validation. We do not claim clinical applicability, and no components of our work should be used for medical diagnosis or decision-making without extensive clinical evaluation.

If any models or code are released, access will be gated under a research-use license, and accompanied by usage guidelines clearly stating that they are intended solely for non-commercial, academic use. The evaluation dataset we construct contains only de-identified medical images drawn from publicly available datasets, and all visual content has been reviewed to ensure it does not pose safety, privacy, or dual-use risks.

O LICENSES

All datasets and models used in this work are publicly available and cited appropriately in the main paper. We do not scrape any new data from the web or repackage any existing datasets; all visual assets have been used in accordance with their licenses.

P USE OF LARGE LANGUAGE MODELS (LLMS)

LLMs were used solely as a writing aid to improve clarity, grammar, and style. They were not involved in generating research ideas, designing methodology, analyzing data, or drawing conclusions.

Q NEW FIGURES DURING REBUTTAL

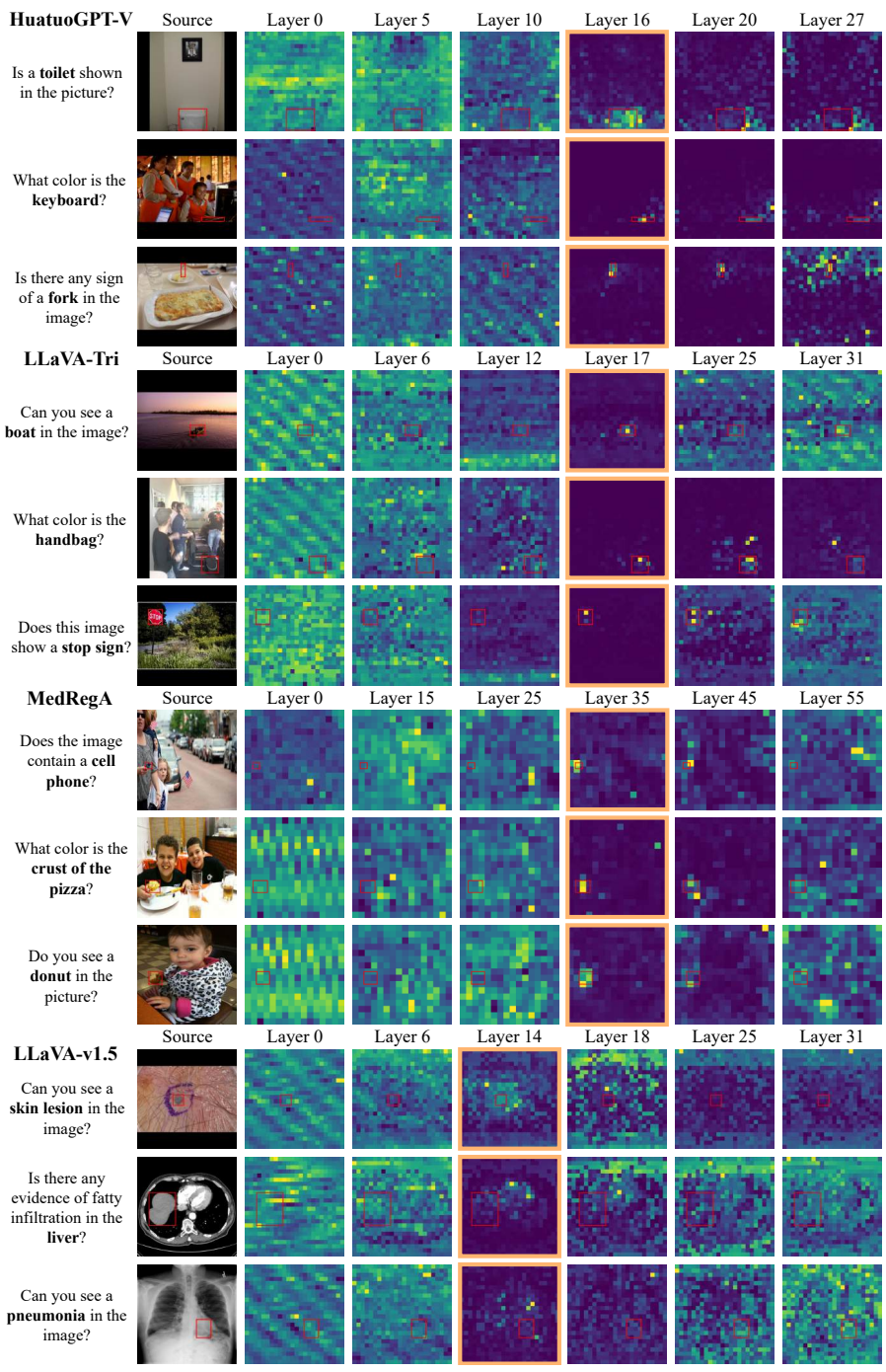

Figure Q.24: **Qualitative** evaluation of (i) medical MLLMs HuatuoGPT-V, LLaVA-Tri and MedRegA on COCO, and (ii) LLaVA-v1.5 on VGMED. We visualize attention maps across different layers, including those with the lowest KL divergence (highlighted with an orange boundary), which are indicative of layers 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.**

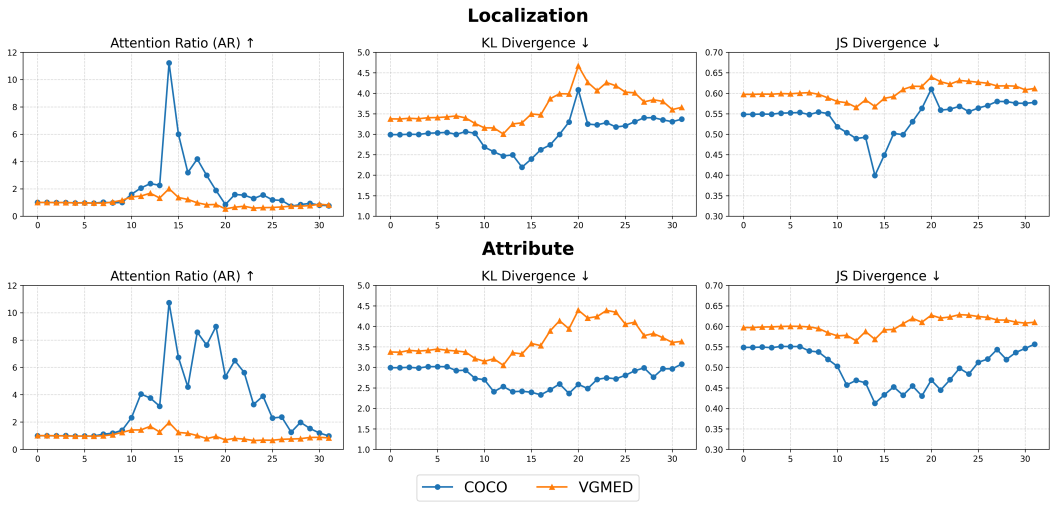

Figure Q.25: **Quantitative** evaluation of LLaVA-v1.5 on VGMED. We observe that LLaVA-v1.5 fails to ground predictions in clinically relevant regions when operating on medical images and medical VQA tasks.

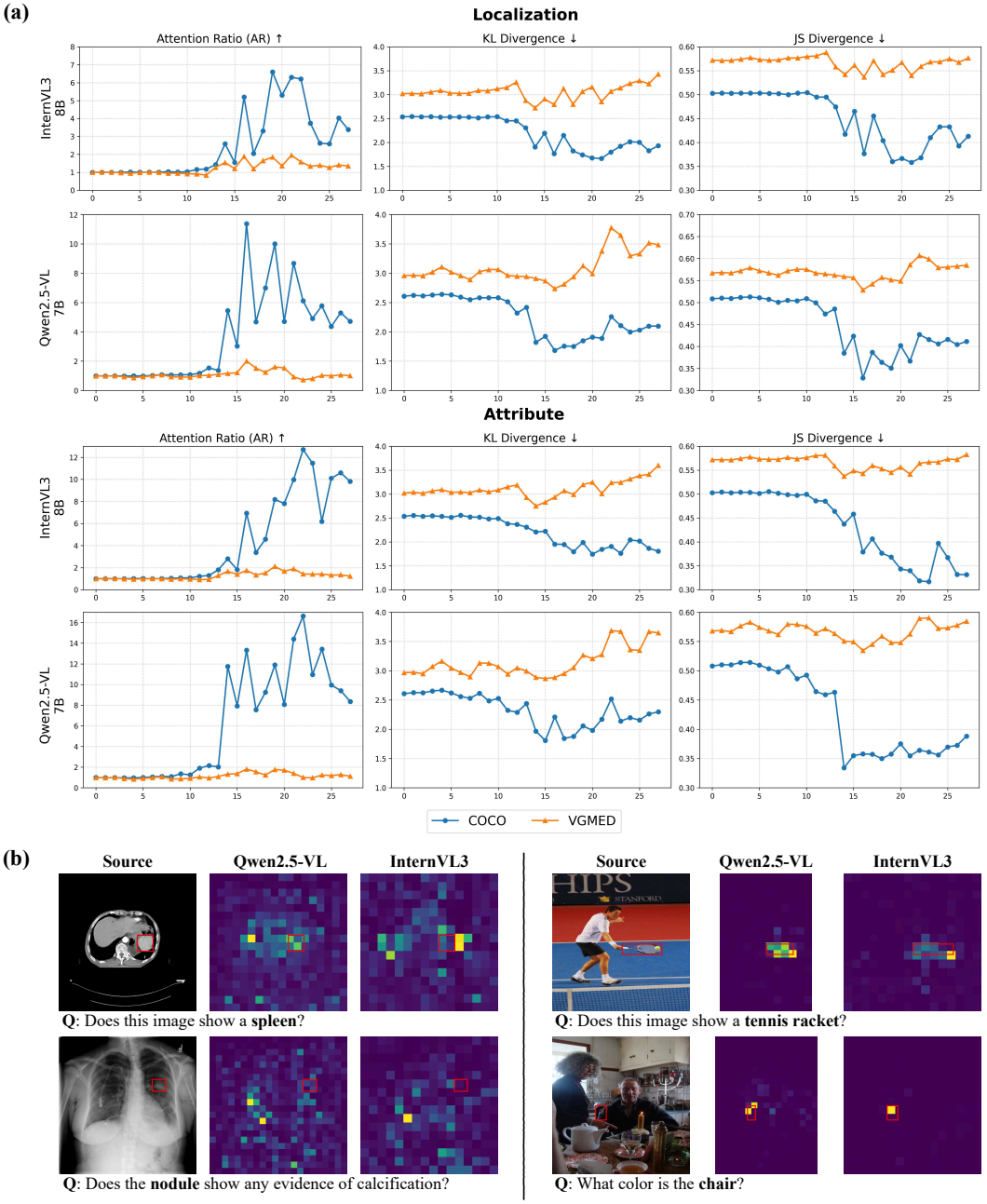

Figure Q.26: (a) **Quantitative** and (b) **qualitative** evaluation of InternVL3-8B and Qwen2.5-VL-7B on VGMED and COCO. We observe that the visual grounding deficiency in medical domain persists even in these latest general-purpose models.

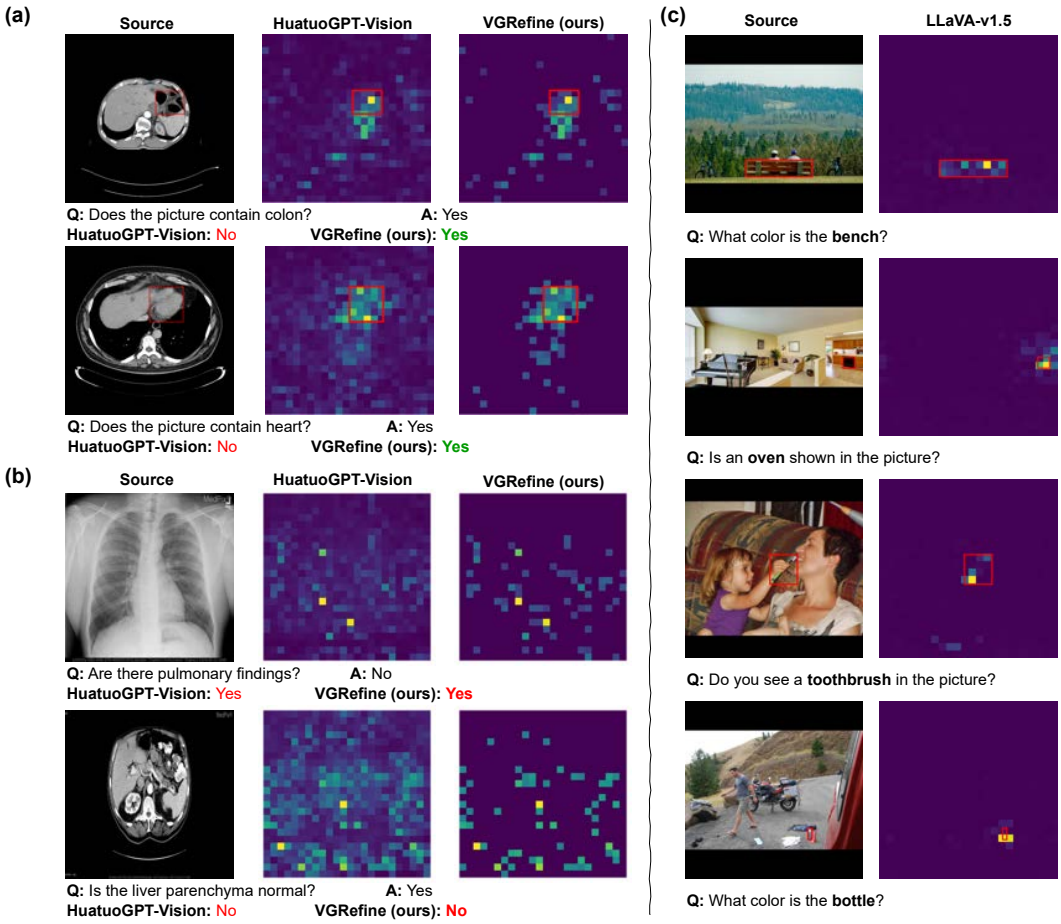

Figure Q.27: **Representative failure cases of HuatuoGPT-Vision on medical benchmarks.** (a) The model correctly interprets the question but attends to the wrong anatomical region, leading to an incorrect answer. After applying VGRefine, the model's attention shifts toward more clinically relevant region, 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.

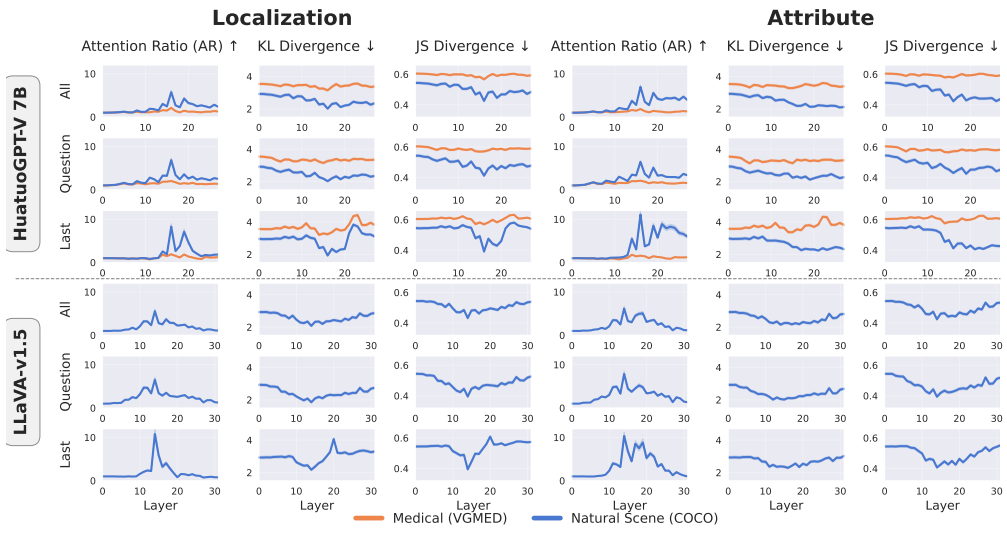

Figure Q.28: **Comparison of visual grounding when using *all input tokens*, *question-only tokens*, or the *last token* to derive attention maps.** Using two representative MLLMs (HuatuoGPT-V-7B and LLaVA-v1.5), we evaluate how different token-selection strategies affect attention alignment on VGMED 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.

R 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. R.29.

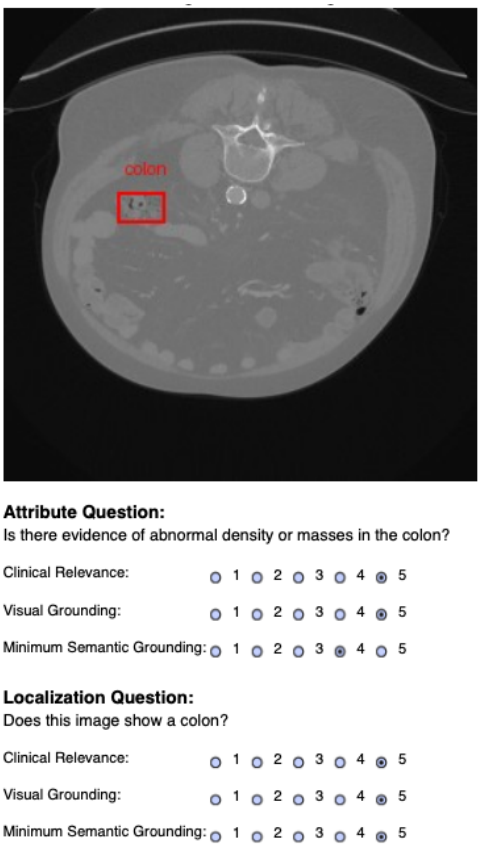

Figure R.29: Example of the clinician rating interface used during VGMED curation.

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. R.6, 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 R.6: 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