# Rethinking Visual Attribution for Chest X-ray Reasoning in Large Vision Language Models

**Guangzhi Xiong**
University of Virginia
guangzhi@virginia.edu

**Qiao Jin**
National Institutes of Health
qiao.jin@nih.gov

**Sanchit Sinha**
University of Virginia
sanchit@virginia.edu

**Zhiyong Lu**
National Institutes of Health
zhiyong.lu@nih.gov

**Aidong Zhang**
University of Virginia
aidong@virginia.edu

## Abstract

Large Vision Language Models (LVLMs) show promise in medical applications, but their inability to faithfully ground responses in visual evidence raises serious concerns about clinical trustworthiness. While visual attribution methods are widely used to explain LVLM predictions, whether these explanations actually reflect the visual evidence underlying the model's decision is largely unverified, since ground-truth annotations for internal model reasoning are typically unavailable. We address this question for chest X-ray (CXR) reasoning by developing a causal evaluation framework that retains only CXR-VQA samples for which the expert-annotated region is verified, via counterfactual editing, to be causally responsible for the model's prediction. Using this framework across 11 attribution methods, six open-source LVLMs, and two output modes (direct answer and step-by-step reasoning), we find that existing attribution methods often fail to identify the evidence used by LVLMs. To address this failure, we propose MedFocus, a concept-based attribution method that localizes clinically meaningful anatomical regions via unbalanced optimal transport and measures their causal effect on model outputs through targeted interventions. MedFocus produces spatial, concept-level, and token-level attributions and substantially outperforms prior methods, taking a step toward more trustworthy attribution for medical LVLMs. Our data and code are available at https://github.com/gzxiong/medfocus/.

## 1 Introduction

Large Vision Language Models (LVLMs) [40, 38] have shown strong capabilities across multimodal tasks such as visual question answering (VQA), captioning, and grounding [37, 7, 76, 43], and are increasingly deployed in medical applications such as radiology report generation [51, 16], medical VQA [78], and diagnostic assistance [16]. As these models are increasingly deployed in high-stakes medical scenarios, a critical concern arises regarding the ability to faithfully attribute the model output to the specific visual evidence in the input. Reliable attribution is essential for clinician trust, error detection, and patient safety, but it remains a largely unsolved challenge for modern LVLMs [12, 57, 71, 27].

Several families of attribution methods have been adapted to LVLMs, including gradient-based saliency [59, 14, 63, 61, 60], attention-based aggregation [72, 4, 1], perturbation-based occlusion [22, 54, 77], and prompting-based grounding [52, 34, 70]. While these approaches offer useful insights, there is a lack of reliable ground truth to objectively evaluate their attribution quality. In practice, determining which visual evidence truly supports the output of a black-box model is

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

inherently challenging, as human annotations can be subjective and may not align with the model's internal reasoning process [5, 20, 30, 33]. This absence of objective evaluation criteria makes it difficult to compare attribution methods rigorously or to identify when they fail, which is particularly dangerous in safety-critical medical applications.

To enable rigorous evaluation of attribution faithfulness, we develop a causal evaluation framework on chest X-ray (CXR) data, the medical modality for which both expert spatial annotations and a region-localized counterfactual editor are publicly available. From three CXR datasets with such annotations [69, 48, 21], we build binary VQA samples and apply a three-step causal filter that retains only those where the annotated region is verified, via counterfactual image editing, to be causally necessary for the model's prediction. The resulting evaluation set, MedGround-Bench, contains 3940 samples across six LVLMs and two output modes. Using it to evaluate 11 widely used attribution methods, we find that none reliably identifies the visual evidence driving LVLM medical predictions, a failure that holds across different settings.

To address this failure, we propose MedFocus, a concept-based causal attribution method for medical LVLM reasoning. Unlike existing post-hoc methods that operate on raw pixel features or internal model representations, MedFocus first segments clinically meaningful regions (e.g., left lung, cardiac silhouette) within the input image, and then evaluates how each region causally influences the model's output. On MedGround-Bench, MedFocus substantially improves over prior methods across all evaluated LVLMs and datasets. By grounding attributions in clinically named concepts, MedFocus produces explanations that are not only more faithful but also directly interpretable by clinicians, bridging low-level visual evidence and high-level clinical understanding. In summary, our contributions are as follows:

- Through a rigorous causal evaluation framework, we show that existing attribution methods consistently fail to faithfully identify the visual evidence underlying medical LVLM predictions. This finding holds across 11 attribution methods, six LVLMs (both generalist and medical), three CXR datasets, and two reasoning modes.
- We propose MedFocus, a concept-based causal attribution method that grounds explanations in clinically meaningful anatomical regions and measures their influence through targeted interventions, producing spatial, concept-level, and token-level attribution outputs that substantially outperform prior methods.
- We release MedGround-Bench, the causally-validated CXR-VQA evaluation suite that enables this study, to support rigorous attribution evaluation in future work.

## 2 Related Work

**Large Vision Language Models (LVLMs) in Medicine.** LVLMs [40, 8, 65] have demonstrated strong capabilities in joint visual and textual understanding, motivating their adaptation to the medical domain in models like LLaVA-Med [36], MedGemma [58], Med-PaLM M [66], and Lingshu [73] for various multimodal medical tasks such as radiology report generation [64, 16], medical visual question answering [25, 35], and diagnostic assistance [47]. While these models achieve impressive performance, their deployment in high-stakes clinical settings has raised growing concerns about trustworthiness and interpretability [49, 62].

**Attribution for Large Vision Language Models.** Existing attribution methods for neural networks fall into four families. Gradient-based methods backpropagate through the network to identify input regions most influencing the output [59, 63, 14]. Attention-based methods aggregate transformer attention weights to highlight attended patches [15, 1]. Perturbation-based methods modify portions of the input and observe how the output changes [77, 54, 42]. Prompting-based approaches ask LVLMs to identify the visual evidence supporting their predictions [52, 34, 70]. Most of these techniques were designed for classification or unimodal settings and transfer poorly to autoregressive multimodal generation.

**Benchmarks for Visual Grounding and Attribution Evaluation.** General-domain grounding benchmarks such as Flickr30k Entities [56] and RefCOCO [46] evaluate a model's ability to localize objects from natural-language descriptions, while medical datasets with radiologist-provided spatial annotations [11, 69, 48, 21] enable analogous phrase-level grounding on clinical images. However, these resources measure localization accuracy against expert annotations rather than whether an

attribution method faithfully identifies the visual evidence driving the model's prediction. In practice, a model may arrive at a correct answer using spurious cues outside the annotated region.

**Causal and Concept-based Interpretability.** Causal interpretability uses counterfactual reasoning to identify input features that drive model predictions, with interventions ranging from simple occlusion [77] to realistic inpainting with editing models [53, 3, 68]. Concept-based interpretability connects low-level features to human-understandable concepts via methods such as TCAV [31], Network Dissection [9], and Concept Bottleneck Models [32, 23, 75]. In medical imaging, anatomical segmentation via atlas-based registration [26], optimal transport [67], or foundation models like MedSAM [44, 45] provides clinically meaningful regions that serve as interpretable concepts for explanation and attribution.

## 3 A Causal Framework for Evaluating CXR Attribution Faithfulness

Evaluating attribution faithfulness requires samples with a causally validated reference region. Starting from CXR VQA data with expert-annotated regions, we filter to retain only samples where the annotated region is verified, via counterfactual editing, to causally drive the model's prediction (Figure 1). The resulting evaluation set, MedGround-Bench, supports attribution analysis across multiple LVLMs and output modes. We focus on CXR because it is currently the only medical modality with both expert spatial annotations and a region-localized counterfactual editing model publicly available, while the construction recipe itself is modality-agnostic.

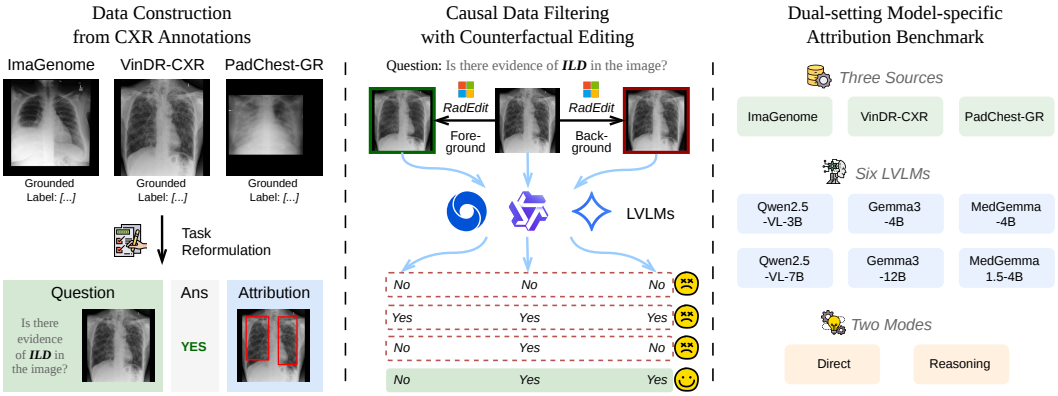


Figure 1: Overview of the construction of MedGround-Bench for CXR attribution evaluation.

### 3.1 Grounded Medical VQA from CXR Annotations

Our framework draws on three publicly available CXR datasets that provide spatially grounded attribute annotations, including ImaGenome [69], VinDR-CXR [48], and PadChest-GR [21]. Each dataset contains radiological images annotated with bounding boxes corresponding to clinically relevant attributes such as diseases or anatomical findings. From these sources, we reformulate the annotated findings as binary VQA samples using a fixed template: *"Is there evidence of* `[attribute]` *in the image?"* This formulation allows for straightforward judgment of model output correctness, which is essential for the subsequent causal filtering steps. For each question, an associated bounding box is provided to indicate the visual evidence identified by human experts. These boxes define the counterfactual edits and, once causally validated, serve as the reference region for evaluation.

### 3.2 Causal Data Filtering with Counterfactual Editing

Since our goal is to evaluate how faithfully attribution methods identify the visual evidence underlying a model's decision, we require samples for which the annotated attribution region is causally linked to the model's output. We apply a three-step filtering process to the constructed VQA data to obtain a high-quality evaluation set.

**Correctness Filtering.** We first query a target LVLM with each VQA question and retain only those questions that the model answers correctly. Questions that are incorrectly answered are discarded, as the ground-truth attribution for an incorrect prediction cannot be reliably established.

**Foreground Counterfactual Editing.** For each remaining question, we generate a counterfactual image by editing the original CXR to remove the target attribute from the annotated region. Specifically, we prompt RadEdit [53] with the union of the finding's annotated boxes as the editing mask, instructing it to inpaint the region such that the attribute is no longer present. We then re-query the model with the same question on the edited image and retain only those samples where the model flips its answer. This ensures that the annotated region is necessary for the model's original prediction.

**Background Counterfactual Editing.** To further reduce noise, we create a second set of counterfactual images by editing the background of the original image, i.e., the region outside the bounding box annotation. We retain only those samples where the model's answer remains unchanged after the background edit. This additional check confirms that the model's decision change in the foreground counterfactual editing is specifically caused by alterations within the annotated region, rather than being an artifact of sensitivity to any image modification.

After all three filtering steps, we obtain a curated evaluation set in which each sample has a verified causal link between the annotated region and the model prediction, providing reliable ground truth for attribution evaluation.

### 3.3 Dataset Statistics and Evaluation Metrics

Our framework supports two output modes, including a direct mode where the model answers yes/no immediately, and a reasoning mode where it produces a step-by-step chain before the final answer. The same causal filtering is applied to both. We focus on six open-source LVLMs spanning generalist and medical families and different scales, including Qwen2.5-VL-3B, Qwen2.5-VL-7B [8], Gemma3-4B, Gemma3-12B [65], MedGemma-4B, and MedGemma1.5-4B [58], since gradient- and attention-based baselines require access to internal hidden states. After filtering, we obtain 1,880 samples for the direct mode (MedGround-Bench-Direct) and 2,060 for the reasoning mode (MedGround-Bench-Reason) across all models and datasets. We measure spatial alignment between predicted attributions and ground-truth bounding boxes using IoU, precision, recall, and F1. Pixel-level saliency maps are converted to bounding boxes via a uniform thresholding procedure. More details about the dataset construction and evaluation can be found in Appendix B.

## 4 MedFocus: Concept-based Causal Attribution for Medical Reasoning

We propose MedFocus, a concept-based attribution method for LVLM medical reasoning outputs. As shown in Figure 2, MedFocus first segments clinically meaningful anatomical regions in the medical image, then measures their causal influence on the model output via targeted interventions. Unlike pixel-level saliency methods, MedFocus produces three complementary forms of attribution. The bounding box of the most causally important region(s) provides a spatial attribution, the name of the attributed anatomy provides a concept-level textual explanation (e.g., "cardiac silhouette"), and for reasoning outputs, the token-level probability changes identify which parts of the reasoning chain are most affected by intervention. While we instantiate MedFocus on CXR using predefined anatomical concepts, the approach is modality-agnostic given suitable concept definitions.

### 4.1 Concept Segmentation via Unbalanced Optimal Transport

We use the 11 anatomical regions predefined in the ImaGenome dataset [69] as our concept vocabulary, including the cardiac silhouette, left/right lung, mediastinum, and other thoracic structures routinely used by radiologists for CXR interpretation. Building on named anatomical concepts aligns attributions with clinical descriptions, and the vocabulary can be extended with finer-grained concepts when reliable masks are available. The full list and a discussion of vocabulary granularity are provided in Appendix D.

**Unbalanced Optimal Transport Mapping.** Given a target CXR image, we localize each anatomical concept by computing an unbalanced optimal transport (UOT) [17, 18] mapping from a reference normal CXR with known anatomical annotations (selected from ImaGenome [69]; details in Appendix

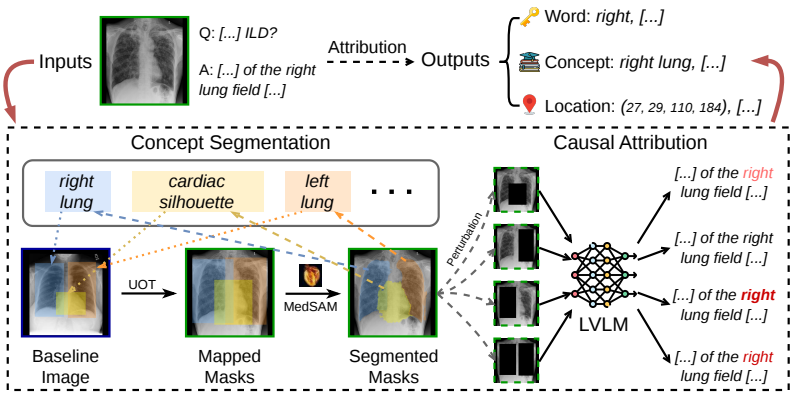


Figure 2: Overview of the proposed MedFocus attribution pipeline. Words significantly affected by the perturbation are highlighted in red.

D) to the target image. We use UOT rather than balanced OT [10, 55] because the mapping between a normal reference and a potentially abnormal target is inherently unbalanced. Pathological changes (e.g., pleural effusion, cardiomegaly) alter local tissue distribution, so the total "mass" of anatomical structures is not conserved, and UOT relaxes the marginal constraints to accommodate this.

Let $\mathbf{x}_{\text{ref}} \in \mathbb{R}^{H\times W}$ denote the reference image with known segmentation masks and $\mathbf{x}_{\text{tgt}} \in \mathbb{R}^{H\times W}$ the target image. We flatten each image into a set of pixel locations and define empirical distributions $\mu_{\text{ref}}$ and $\mu_{\text{tgt}}$ weighted by normalized intensity $\mu_{\text{ref}}(i) = x_{\text{ref}}^{(i)} / \sum_k x_{\text{ref}}^{(k)}$ (analogously for $\mu_{\text{tgt}}$). The transport cost $C_{ij}$ is the squared Euclidean distance between the spatial coordinates of pixel $i$ in $\mathbf{x}_{\text{ref}}$ and pixel $j$ in $\mathbf{x}_{\text{tgt}}$. We then solve for the UOT plan $\mathbf{T}^*$:

$$\mathbf{T}^* = \arg\min_{\mathbf{T}\geq 0} \sum_{i,j} C_{ij}\, T_{ij} + \lambda_1\, D_{\text{KL}}(\mathbf{T}\mathbf{1} \,\|\, \mu_{\text{ref}}) + \lambda_2\, D_{\text{KL}}\big(\mathbf{T}^\top \mathbf{1} \,\|\, \mu_{\text{tgt}}\big)\,, \tag{1}$$

where $D_{\text{KL}}$ is the KL divergence and $\lambda_1, \lambda_2 > 0$ control marginal relaxation. For each concept $c$ with reference pixel set $\mathcal{S}_c^{\text{ref}}$, we transport its mass through $T^*$ to obtain the corresponding target region $\mathcal{S}_c^{\text{tgt}}$, with more details available in Appendix D.

**Mask Refinement with MedSAM.** Since UOT-derived pixel sets may have noisy boundaries, we refine each transferred region using MedSAM [44]. For each concept $c$, we compute the tightest bounding box enclosing $\mathcal{S}_c^{\text{tgt}}$ and use it as a box prompt to MedSAM, producing a clean mask $\mathbf{M}_c \in \{0,1\}^{H\times W}$. The effectiveness of this refinement step is validated in Section 5.4.

### 4.2 Causal Attribution via Concept Intervention

Given concept masks, we attribute model predictions by intervening on each concept and measuring the resulting change in output.

**Counterfactual Generation.** For concept $c$ with mask $\mathbf{M}_c$, we generate a counterfactual by zero-masking its bounding box:

$$\tilde{\mathbf{x}}_c = \mathbf{x}_{\text{tgt}} \odot (\mathbf{1} - \mathbf{B}_c), \tag{2}$$

where $\mathbf{B}_c \in \{0,1\}^{H\times W}$ is the bounding box mask and $\odot$ denotes element-wise multiplication. Using the bounding box rather than the pixel-level mask ensures sufficient contextual removal for a cleaner causal signal. Ablations in Section 5.4 confirm that bounding box masking provides a stronger attribution signal than pixel-level masking or generative counterfactual editing.

**Measuring Output Change.** Let $\mathbf{y} = (y_1, \ldots, y_T)$ denote the model's original output given image $\mathbf{x}_{\text{tgt}}$ and question $q$. Rather than regenerating the full output for each counterfactual, we run a single forward pass on $\tilde{\mathbf{x}}_c$ conditioned on $\mathbf{y}$ and measure the cumulative drop in token-level log-probabilities:

$$\Delta_c = \sum_{t=1}^{T} \max\big(0,\ \log p(y_t \mid \mathbf{x}_{\text{tgt}}, q, \mathbf{y}_{<t}) - \log p(y_t \mid \tilde{\mathbf{x}}_c, q, \mathbf{y}_{<t})\big)\,, \tag{3}$$

where a larger $\Delta_c$ implies a stronger causal contribution of concept $c$. Conditioning on the original sequence $y$ rather than regenerating isolates each concept's effect on the prediction the model actually produced, avoids sampling noise, and requires only one forward pass per concept. The $\max(0, \cdot)$ operator restricts attribution to probability drops, since an increase upon removing a region reflects a contradictory rather than supporting cue.

**Composite Concept Attribution.** The model's prediction may rely on multiple anatomical regions jointly. For a clinically meaningful composite group $\mathcal{C}'$ (e.g., left and right lungs combined), we additionally evaluate:

$$\tilde{\mathbf{x}}_{\mathcal{C}'} = \mathbf{x}_{\text{tgt}} \odot \left( \mathbf{1} - \bigcup_{c \in \mathcal{C}'} \mathbf{B}_c \right), \tag{4}$$

with $\Delta_{\mathcal{C}'}$ computed analogously. The set of composite groups $\mathcal{G}$ is predetermined based on clinical relevance, keeping the method efficient.

**Attribution Output.** The concept (or composite group) inducing the largest output change is identified as most causally relevant:

$$c^* = \arg\max_{c \in \mathcal{C} \cup \mathcal{G}} \Delta_c. \tag{5}$$

In practice, this scoring does not simply favor the largest mask. MedFocus can select localized evidence rather than broader regions, as shown in Figure 4. The bounding box of $c^*$ is reported as the spatial attribution (directly comparable with ground-truth annotations in our evaluation), the name of $c^*$ as the concept-level explanation, and the per-token contributions to $\Delta_{c^*}$ over reasoning outputs as the token-level attribution.

**Concept Relevance Thresholding.** Since LVLM reasoning can be complex and noisy, the prediction may not rely on any predefined anatomical concept. We detect such cases via a threshold on the relative probability ratio $r_c = \exp(-\Delta_c)$. If $\min_{c \in \mathcal{C} \cup \mathcal{G}} r_c \geq \tau$ (we use $\tau = 0.75$), we conclude that no single concept drives the prediction and default to using the entire image as the attribution result.

# 5 Experiments

## 5.1 Attribution Evaluation with MedGround-Bench

We use MedGround-Bench to evaluate the faithfulness of 11 existing attribution methods spanning attention-based methods [1, 6], gradient-based methods [15, 59, 14, 63], prompting-based pipelines [44], and perturbation-based approaches [77, 54], alongside our proposed MedFocus. All methods are evaluated using Intersection over Union (IoU), F1 score (F1), Precision (Prec), and Recall. Implementation details for baselines and MedFocus are provided in Appendices C and D.

Table 1: Comparison of visual attribution methods on MedGround-Bench-Direct. All scores are percentages. Best results are in **bold** and second best are underlined.

| Method | ImaGenome | | | | VinDR-CXR | | | | PadChest-GR | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | IoU | F1 | Prec | Recall | IoU | F1 | Prec | Recall | IoU | F1 | Prec | Recall |
| ***Attention-based Methods*** | | | | | | | | | | | | |
| Attention Head | 15.87 | 26.10 | 48.17 | 19.12 | 6.93 | 11.80 | 11.09 | 23.53 | 14.33 | 23.30 | 31.42 | 23.53 |
| Attention Rollout [1] | 2.70 | 5.05 | 12.70 | 3.30 | 0.70 | 1.30 | 1.58 | 1.67 | 2.77 | 5.06 | 9.29 | 4.17 |
| LRP [6] | 5.67 | 10.15 | 22.61 | 6.84 | 1.93 | 3.42 | 3.39 | 7.34 | 4.12 | 7.18 | 11.79 | 6.89 |
| ***Gradient-based Methods*** | | | | | | | | | | | | |
| Gradient-weighted Attn [15] | 39.24 | 54.80 | 39.25 | **99.90** | 7.73 | 13.26 | 7.73 | **100.00** | 22.73 | 34.21 | 22.73 | **100.00** |
| GradCAM [59] | 34.47 | 49.10 | 44.81 | 80.62 | 8.53 | 14.36 | 11.95 | 68.91 | 20.33 | 31.22 | 26.01 | 78.36 |
| GradCAM++ [14] | 30.54 | 44.07 | 44.40 | 65.42 | 7.40 | 12.61 | 9.36 | 60.64 | 18.28 | 28.09 | 25.93 | 62.96 |
| Integrated Gradients [63] | 11.71 | 19.13 | 46.39 | 13.45 | 9.38 | 14.96 | 15.10 | 30.42 | 13.06 | 20.12 | 33.85 | 20.73 |
| ***Prompting-based Methods*** | | | | | | | | | | | | |
| Prompting | 8.24 | 12.17 | 15.24 | 17.73 | 2.45 | 4.04 | 3.24 | 12.22 | 7.55 | 11.47 | 13.03 | 18.02 |
| Prompting + MedSAM [44] | 37.62 | 50.56 | 46.62 | 74.64 | 8.33 | 13.78 | 9.11 | 86.08 | 21.76 | 32.29 | 23.22 | 85.52 |
| ***Perturbation-based Methods*** | | | | | | | | | | | | |
| Occlusion [77] | 22.16 | 33.48 | 60.25 | 36.72 | 13.62 | 21.28 | **22.13** | 43.81 | 20.56 | 31.20 | **44.03** | 40.72 |
| RISE [54] | 19.17 | 30.84 | 50.35 | 24.18 | 10.14 | 16.69 | 14.21 | 36.69 | 16.80 | 26.89 | 33.76 | 29.45 |
| **Ours (MedFocus)** | **54.28** | **67.52** | **64.65** | 80.06 | **14.86** | **23.10** | 16.00 | 79.86 | **33.02** | **45.69** | 41.05 | 73.25 |

Table 1 presents the comparison on MedGround-Bench-Direct, with metrics averaged across all models. No existing attribution method achieves consistently faithful attribution on this benchmark. Even on samples filtered to have a verified causal link between the annotated region and the model prediction, baselines either produce diffuse maps with low precision or focused maps that miss the true evidence. Attribution methods such as GradCAM and Integrated Gradients, which are used frequently for visual classifiers, perform poorly in the LVLM setting. While some baselines (e.g., Gradient-weighted Attention) achieve near-perfect recall, they suffer from very low precision, indicating overly broad highlighted regions. In contrast, MedFocus consistently achieves the best IoU and F1 across all three datasets, maintaining a strong precision-recall balance while accurately localizing diagnostically relevant regions.

In Table 1, all saliency maps are converted to boxes with a single fixed threshold (Appendix C). To test whether this disadvantages the baselines, we tune the threshold separately for each baseline and LVLM on the 144 held-out PadChest-GR tuning questions (Appendix B.1). As shown in Table 2, tuning improves the IoU and F1 of the baselines, mainly by trading precision for recall, but they remain well below MedFocus, which requires no threshold.

Table 2: Attribution performance on MedGround-Bench-Direct with fixed and validation-tuned saliency thresholds, averaged across all datasets and models.

| **Method** | **Threshold** | **IoU** | **F1** | **Prec** | **Recall** |
|---|---|---|---|---|---|
| Attention Head | Fixed | 14.01 | 22.98 | 35.32 | 22.56 |
| | Tuned | 25.29 | 37.55 | 30.77 | 70.18 |
| Integrated Gradients [63] | Fixed | 13.24 | 20.69 | 37.14 | 21.62 |
| | Tuned | 26.40 | 38.47 | 34.54 | 75.53 |
| Occlusion [77] | Fixed | 20.41 | 30.96 | 44.90 | 41.42 |
| | Tuned | 25.84 | 37.92 | 37.57 | 67.12 |
| **Ours (MedFocus)** | – | **37.93** | **49.82** | **45.37** | **77.68** |

Figure 3 shows the evaluation results on the reasoning set, where the attribution target is set as the probability of the whole generated sequence for all methods to enable fair comparison. Consistent with the direct results, existing methods fail to faithfully attribute reasoning outputs, with many showing substantial performance drops (e.g., GradCAM++ drops from 30.54% to 23.70% IoU on ImaGenome). MedFocus maintains strong attribution quality (e.g., 52.95% IoU on ImaGenome), as its causal attribution framework avoids probing model internals and is robust to multi-step reasoning.

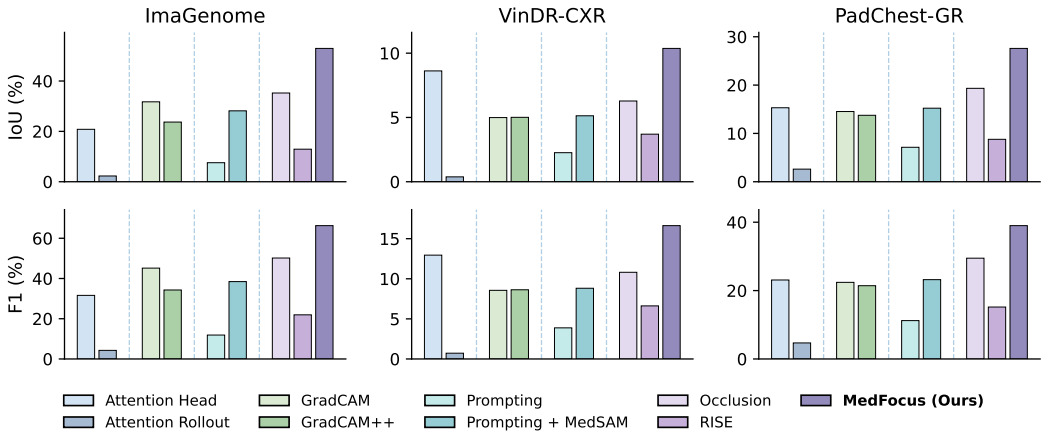


Figure 3: Reasoning attribution evaluation on MedGround-Bench-Reason. Metrics are averaged across all models.

### 5.2 Qualitative Analysis of Attribution Quality

Beyond evaluation metrics, qualitative examples reveal clear differences in how attribution methods localize the evidence underlying LVLM predictions. Figure 4 compares representative cases from the three source datasets, including lobar / segmental collapse from ImaGenome, interstitial lung disease (ILD) from VinDR-CXR, and cardiomegaly from PadChest-GR. Across all three examples, existing baselines often produce either diffuse attributions that cover large portions of the image or misplaced regions that only weakly overlap with the annotated evidence. In contrast, MedFocus produces tighter and more clinically plausible localizations, with predicted regions that align more closely with the ground-truth boxes across diverse findings and datasets.

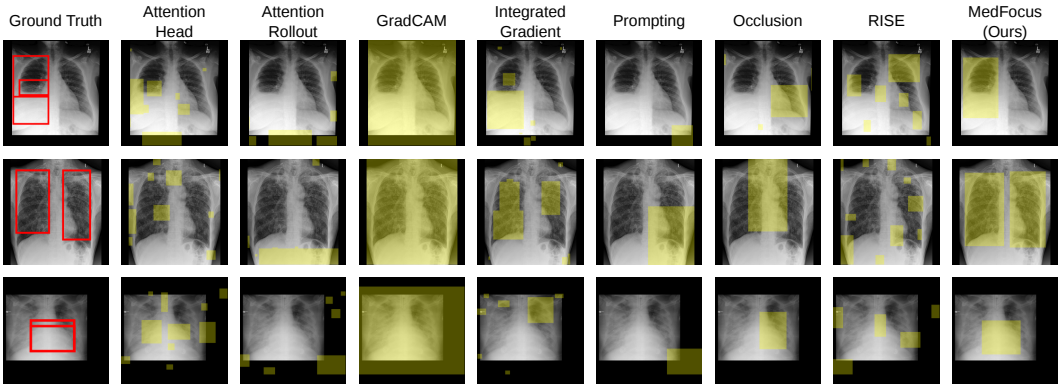


Figure 4: Qualitative comparison on three MedGround-Bench-Direct examples. Ground-truth evidence is shown in red and predicted attributions are in yellow.

Figure 5 further illustrates the advantage of MedFocus in the reasoning setting. Instead of attributing the entire reasoning chain to a single diffuse heatmap, MedFocus can track which anatomical concepts support different parts of the generated rationale. In the illustrated example, earlier tokens are associated with broad lung-level context, whereas later and more clinically specific phrases become concentrated on the cardiac silhouette region, consistent with the final diagnosis. This progressive refinement suggests that MedFocus captures not only where the model looks, but also how visual evidence is recruited over the course of multi-step reasoning.

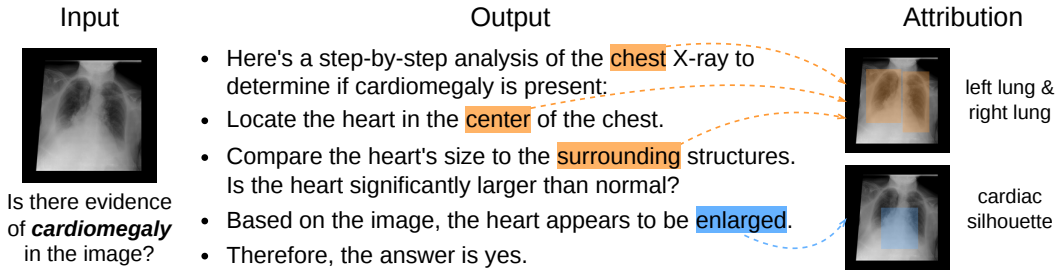


Figure 5: Token-level concept attribution for a MedGround-Bench-Reason example.

### 5.3 LVLM Attribution across Models and Sample Groups

Figure 6 compares MedFocus attribution across three progressively filtered sample groups from the MedGround-Bench construction pipeline: G1 (incorrectly answered samples removed by correctness filtering), G2 (correct but ungrounded samples removed by causal filtering), and G3 (samples retained in MedGround-Bench). Although annotated regions in G1 and G2 may not reflect the model's actual reasoning, we compute their IoU with MedFocus attributions to assess how detected model evidence aligns with human annotations across groups. We also report the failure rate, i.e., the proportion of samples where the model does not use any anatomical concept, as defined in Section 4.2.

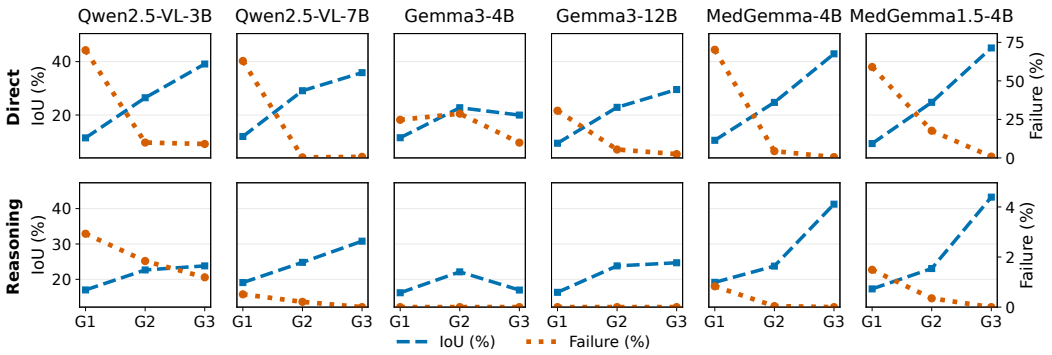


Figure 6: Comparison of MedFocus attributions across models and sample groups.

From Figure 6, we observe a consistent pattern across models that the IoU score improves from G1 to G3. In both the direct and reasoning settings, IoU generally increases as the samples become more causally grounded, while the failure rate decreases. This trend indicates that the MedGround-Bench filtering pipeline effectively removes cases in which models either rely on irrelevant visual cues or produce correct answers for the wrong reasons, leaving a final set whose predictions are more cleanly tied to clinically meaningful evidence. Another notable trend is that failure rates in the reasoning mode are substantially lower than in the direct-answer mode for all models, often approaching zero on G3. This suggests that generating intermediate reasoning steps encourages LVLMs to engage more consistently with anatomically meaningful evidence, even when the final prediction is still incorrect or only partially grounded.

Comparing models, the medically trained models, MedGemma1.5-4B and MedGemma-4B, exhibit the strongest attribution behavior on G3, with higher IoU and lower failure rates than the generalist Qwen2.5-VL and Gemma3 models, especially in the reasoning setting. Table 5 also shows that medically trained models have a larger proportion of correct-and-grounded samples after causal filtering. Within the same model family, larger models also tend to show better G3 attribution in reasoning mode, suggesting that increased model capacity improves the alignment between generated reasoning and visual evidence. Overall, these results indicate that both domain-specific medical training and larger model scale improve the faithfulness of visual grounding, while smaller general-purpose models remain harder to attribute reliably. We further evaluate Lingshu-7B [73], a medical LVLM fine-tuned from Qwen2.5-VL-7B, in the direct-answer mode. Lingshu-7B yields about twice the share of correct and grounded samples of its base model (4.9% vs. 2.3%), and MedFocus again achieves the best IoU and F1 on all three datasets (Appendix E.6).

### 5.4 Ablation Studies

We ablate three key design dimensions in our MedFocus method, namely the segmentation paradigm, the localization strategy within the two-stage segmentation pipeline, and the counterfactual intervention strategy for causal attribution. The results are summarized in Table 3.

**Segmentation paradigm.** Compared with end-to-end segmentation variants [13, 39, 50], our two-stage design, which combines UOT-based localization with MedSAM refinement, consistently yields higher IoU and F1 scores. Although medical segmentation models such as MedSAM3 [39] and RadZero [50] achieve relatively high precision, their substantially lower recall leads to worse attribution quality.

**Localization strategy.** Within the two-stage framework, we vary the localization method and whether MedSAM refinement is applied. Table 3 shows that Grounding-DINO-based localization [41] provides extremely high recall but low precision, suggesting overly broad boxes that dilute attribution specificity. In contrast, UOT provides a better precision-recall balance and achieves higher IoU and F1. Our ablation further shows that MedSAM refinement improves attribution quality on top of UOT-based localization.

**Counterfactual intervention strategy.** For the counterfactual intervention strategy, we vary both the region removed during intervention (segmentation mask vs. bounding box) and the counterfactual

generation method (RadEdit inpainting vs. zero masking). The best-performing variant combines bounding-box intervention with zero masking, outperforming mask-based intervention and improving over RadEdit-based counterfactuals. A further comparison with cross-disease replacement, which uses RadEdit to depict a different finding in the concept region, confirms that zero masking yields the best IoU and F1 (Appendix E.5). Moreover, zero masking the concept selected by MedFocus lowers the probability of the original answer far more than masking a same-size region that does not overlap it (Appendix E.5), indicating that the effect is concept-specific rather than an artifact of masking. These findings validate the design choices in Section 4.2.

Table 3: Ablation study on three design dimensions of MedFocus: the segmentation paradigm, the localization strategy, and the counterfactual intervention strategy.

| **Ablation** | **Method** | **IoU ↑** | **F1 ↑** | **Prec ↑** | **Recall ↑** |
|---|---|---|---|---|---|
| *Paradigm* | SAM3 (end-to-end) | 30.89 | 42.61 | 38.75 | 79.11 |
| | MedSAM3 (end-to-end) | 33.52 | 45.02 | 53.19 | 64.04 |
| | RadZero (end-to-end) | 6.95 | 12.13 | 59.68 | 9.29 |
| | **UOT + MedSAM (detect+seg)** | **37.93** | **49.82** | 45.37 | 77.68 |
| *Localization* | Grounding DINO | 27.72 | 39.63 | 27.74 | 99.77 |
| | Grounding DINO + MedSAM | 29.96 | 41.82 | 30.09 | 99.25 |
| | UOT | 36.24 | 48.16 | 45.72 | 71.09 |
| | **UOT + MedSAM** | **37.93** | **49.82** | 45.37 | 77.68 |
| *Intervention* | Segmentation mask + RadEdit | 32.66 | 44.40 | 45.15 | 65.00 |
| | Segmentation mask + Zero masking | 33.76 | 45.28 | 43.60 | 69.70 |
| | Bounding box + RadEdit | 32.27 | 43.87 | 44.54 | 64.78 |
| | **Bounding box + Zero masking** | **37.93** | **49.82** | 45.37 | 77.68 |

## 6 Conclusion

This work presents a causal framework for evaluating visual attribution faithfulness in chest X-ray reasoning with LVLMs. Using MedGround-Bench, a causally validated attribution benchmark, we show that existing attention-, gradient-, prompting-, and perturbation-based methods often fail to identify the visual evidence driving model predictions. We then introduce MedFocus, a concept-based causal attribution method that grounds explanations in clinically meaningful anatomical regions and measures their influence through targeted interventions. Across multiple LVLMs, datasets, and output modes, MedFocus yields more faithful and interpretable spatial, concept-level, and token-level attributions, offering a step toward more trustworthy medical LVLM reasoning.

## Acknowledgments and Disclosure of Funding

This research was partly supported by the Intramural Research Program of the National Institutes of Health (NIH). The contributions of the NIH author(s) are considered Works of the United States Government. This research was also partially supported by the US National Science Foundation (NSF) and the NIH under grants 2616033, 2538206, 2529378, 2217071, 2213700, R01LM014012-01A1, 1R01HL184168-01, and the NIH Pathway to Independence Award K99LM014903 (Q.J.). The findings and conclusions presented in this paper are those of the author(s) and do not necessarily reflect the views of the NIH, the NSF, or the U.S. Department of Health and Human Services.

## References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.385. URL https://aclanthology.org/2020.acl-main.385/.

[2] Suhyun Ahn, Wonjung Park, Jihoon Cho, and Jinah Park. Volumetric conditioning module to control pretrained diffusion models for 3d medical images. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 85–95, 2025. doi: 10.1109/WACV61041.2025.00019.

[3] Malek Ben Alaya, Daniel M. Lang, Benedikt Wiestler, Julia A. Schnabel, and Cosmin I. Bercea. Mededit: Counterfactual diffusion-based image editing on brain mri. In Virginia Fernandez, Jelmer M. Wolterink, David Wiesner, Samuel Remedios, Lianrui Zuo, and Adrià Casamitjana, editors, *Simulation and Synthesis in Medical Imaging*, pages 167–176, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73281-2.

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[5] Lora Aroyo and Chris Welty. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24, Mar. 2015. doi: 10.1609/aimag.v36i1.2564. URL https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2564.

[6] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):1–46, 07 2015. doi: 10.1371/journal.pone.0130140. URL https://doi.org/10.1371/journal.pone.0130140.

[7] 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, 2023. URL https://arxiv.org/abs/2308.12966.

[8] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

[9] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[10] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015. doi: 10.1137/141000439. URL https://doi.org/10.1137/141000439.

[11] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. Making the most of text semantics to improve biomedical vision–language processing. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 1–21, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-20059-5.

[12] Katarzyna Borys, Yasmin Alyssa Schmitt, Meike Nauta, Christin Seifert, Nicole Krämer, Christoph M. Friedrich, and Felix Nensa. Explainable ai in medical imaging: An overview for clinical practitioners – beyond saliency-based xai approaches. *European Journal of Radiology*, 162:110786, 2023. ISSN 0720-048X. doi: https://doi.org/10.1016/j.ejrad.2023.110786. URL https://www.sciencedirect.com/science/article/pii/S0720048X23001006.

[13] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan

Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2025. URL https://arxiv.org/abs/2511.16719.

[14] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847, 2018. doi: 10.1109/WACV.2018.00097.

[15] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 782–791, June 2021.

[16] Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, Emily Tsai, Andrew Johnston, Cameron Olsen, Tanishq Mathew Abraham, Sergios Gatidis, Akshay S Chaudhari, and Curtis Langlotz. Chexagent: Towards a foundation model for chest x-ray interpretation. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*, 2024. URL https://openreview.net/forum?id=P3LOmrZWGR.

[17] Lenaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems, 2018. URL https://doi.org/10.1090/mcom/3303.

[18] Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: Dynamic and kantorovich formulations. *Journal of Functional Analysis*, 274 (11):3090–3123, 2018. ISSN 0022-1236. doi: https://doi.org/10.1016/j.jfa.2018.03.008. URL https://www.sciencedirect.com/science/article/pii/S0022123618301058.

[19] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf.

[20] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100, 2017. ISSN 1077-3142. doi: https://doi.org/10.1016/j.cviu.2017.10.001. URL https://www.sciencedirect.com/science/article/pii/S1077314217301649. Language in Vision.

[21] Daniel Coelho de Castro, Aurelia Bustos, Shruthi Bannur, Stephanie L. Hyland, Kenza Bouzid, Maria Teodora Wetscherek, Maria Dolores Sánchez-Valverde, Lara Jaques-Pérez, Lourdes Pérez-Rodríguez, Kenji Takeda, José María Salinas-Serrano, Javier Alvarez-Valle, Joaquín Galant-Herrero, and Antonio Pertusa. Padchest-gr: A bilingual chest x-ray dataset for grounded radiology report generation. *NEJM AI*, 2(7):AIdbp2401120, 2025. doi: 10.1056/AIdbp2401120. URL https://ai.nejm.org/doi/full/10.1056/AIdbp2401120.

[22] Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[23] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/77d2afcb31f6493e350fca61764efb9a-Paper.pdf.

[24] Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. Physiobank, physiotoolkit, and physionet. *Circulation*, 101(23):e215–e220, 2000. doi: 10.1161/01.CIR.101.23.e215. URL https://www.ahajournals.org/doi/abs/10.1161/01.CIR.101.23.e215.

[25] 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.

[26] Juan Eugenio Iglesias and Mert R. Sabuncu. Multi-atlas segmentation of biomedical images: A survey. *Medical Image Analysis*, 24(1):205–219, 2015. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2015.06.012. URL https://www.sciencedirect.com/science/article/pii/S1361841515000997.

[27] Qiao Jin, Yin Fang, Lauren He, Yifan Yang, Guangzhi Xiong, Zhizheng Wang, Nicholas Wan, Joey Chan, Donald C. Comeau, Robert Leaman, Charalampos S. Floudas, Aidong Zhang, Michael F. Chiang, Yifan Peng, and Zhiyong Lu. Med-v1: Small language models for zero-shot and scalable biomedical evidence attribution, 2026. URL https://arxiv.org/abs/2603.05308.

[28] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

[29] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

[30] Ashkan Khakzar, Pedram Khorsandi, Rozhin Nobahari, and Nassir Navab. Do explanations explain? model knows best. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10244–10253, June 2022.

[31] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2668–2677. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kim18d.html.

[32] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5338–5348. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

[33] Satyapriya Krishna, Tessa Han, Alex Gu, Steven Wu, Shahin Jabbari, and Himabindu Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=jESY2WTZCe.

[34] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9579–9589, June 2024.

[35] 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):180251, 2018.

[36] 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, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 28541–28564. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/5abcdf8ecdcacba028c6662789194572-Paper-Datasets_and_Benchmarks.pdf.

[37] 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, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/li23q.html.

[38] Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. A survey of state of the art large vision language models: Benchmark evaluations and challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1587–1606, June 2025.

[39] Anglin Liu, Rundong Xue, Xu R. Cao, Yifan Shen, Yi Lu, Xiang Li, Qianqian Chen, and Jintai Chen. Medsam3: Delving into segment anything with medical concepts, 2025. URL https://arxiv.org/abs/2511.19046.

[40] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.

[41] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 38–55, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72970-6.

[42] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

[43] Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized visual tokenization for grounding multimodal large language models. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 417–435, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72658-3.

[44] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature communications*, 15(1):654, 2024.

[45] Jun Ma, Zongxin Yang, Sumin Kim, Bihui Chen, Mohammed Baharoon, Adibvafa Fallahpour, Reza Asakereh, Hongwei Lyu, and Bo Wang. Medsam2: Segment anything in 3d medical images and videos, 2025. URL https://arxiv.org/abs/2504.03600.

[46] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[47] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

[48] Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, Dung D Le, Chi M Pham, Hang TT Tong, Diep H Dinh, et al. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Scientific Data*, 9(1):429, 2022.

[49] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems, 2023. URL https://arxiv.org/abs/2303.13375.

[50] Jonggwon Park, Byungmu Yoon, Soobum Kim, and Kyoyun Choi. Radzero: Similarity-based cross-attention for explainable vision-language alignment in chest x-ray with zero-shot multi-task capability. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=WQq5JPGQ0C.

[51] Chantal Pellegrini, Ege Özsoy, Benjamin Busam, Benedikt Wiestler, Nassir Navab, and Matthias Keicher. Radialog: Large vision-language models for x-ray reporting and dialog-driven assistance. In *Medical Imaging with Deep Learning*, 2025. URL https://openreview.net/forum?id=trUvr1gSNI.

[52] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. Grounding multimodal large language models to the world. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=lLmqxkfSIw.

[53] Fernando Pérez-García, Sam Bond-Taylor, Pedro P. Sanchez, Boris van Breugel, Daniel C. Castro, Harshita Sharma, Valentina Salvatelli, Maria T. A. Wetscherek, Hannah Richardson, Matthew P. Lungren, Aditya Nori, Javier Alvarez-Valle, Ozan Oktay, and Maximilian Ilse. Radedit: Stress-testing biomedical vision models via diffusion image editing. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 358–376, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73254-6.

[54] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: randomized input sampling for explanation of black-box models. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, page 151. BMVA Press, 2018. URL http://bmvc2018.org/contents/papers/1064.pdf.

[55] Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[56] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[57] Rikard Rosenbacke, Åsa Melhus, Martin McKee, and David Stuckler. How explainable artificial intelligence can increase or decrease clinicians' trust in ai applications in health care: Systematic review. *JMIR AI*, 3:e53207, Oct 2024. ISSN 2817-1705. doi: 10.2196/53207. URL https://ai.jmir.org/2024/1/e53207.

[58] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, Lu Yang, Kejia Chen, Per Bjornsson, Shashir Reddy, Ryan Brush, Kenneth Philbrick, Mercy Asiedu, Ines Mezerreg, Howard Hu, Howard Yang, Richa Tiwari, Sunny Jansen, Preeti Singh, Yun Liu, Shekoofeh Azizi, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Riviere, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Elena Buchatskaya, Jean-Baptiste Alayrac, Dmitry Lepikhin, Vlad Feinberg, Sebastian Borgeaud, Alek Andreev, Cassidy Hardin, Robert Dadashi, Léonard Hussenot, Armand Joulin, Olivier Bachem, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Clement Farabet, Joelle Barral, Tris Warkentin, Jonathon Shlens, David Fleet, Victor Cotruta, Omar Sanseviero, Gus Martins, Phoebe Kirk, Anand Rao, Shravya Shetty, David F. Steiner, Can Kirmizibayrak, Rory Pilgrim, Daniel Golden, and Lin Yang. Medgemma technical report, 2025. URL https://arxiv.org/abs/2507.05201.

[59] 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 (ICCV)*, Oct 2017.

[60] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3145–3153. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/shrikumar17a.html.

[61] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014. URL https://arxiv.org/abs/1312.6034.

[62] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[63] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

[64] Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7433–7442, June 2023.

[65] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud,

Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

[66] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, Anil Palepu, Basil Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David Fleet, Philip Mansfield, Sushant Prakash, Renee Wong, Sunny Virmani, Christopher Semturs, S. Sara Mahdavi, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Karan Singhal, Pete Florence, Alan Karthikesalingam, and Vivek Natarajan. Towards generalist biomedical ai. *NEJM AI*, 1(3):AIoa2300138, 2024. doi: 10.1056/AIoa2300138. URL https://ai.nejm.org/doi/full/10.1056/AIoa2300138.

[67] Wei Wang, John A. Ozolek, Dejan Slepčev, Ann B. Lee, Cheng Chen, and Gustavo K. Rohde. An optimal transportation approach for nuclear structure-based pathology. *IEEE Transactions on Medical Imaging*, 30(3):621–631, 2011. doi: 10.1109/TMI.2010.2089693.

[68] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ze Liu, Ziyi Xia, Chaofan Li, Haoge Deng, Jiahao Wang, Kun Luo, Bo Zhang, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. Omnigen2: Exploration to advanced multimodal generation, 2025. URL https://arxiv.org/abs/2506.18871.

[69] Joy T Wu, Nkechinyere Nneka Agu, Ismini Lourentzou, Arjun Sharma, Joseph Alexander Paguio, Jasper Seth Yao, Edward Christopher Dee, William G Mitchell, Satyananda Kashyap, Andrea Giovannini, Leo Anthony Celi, and Mehdi Moradi. Chest imagenome dataset for clinical reasoning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=H-d5634yVi.

[70] Qiong Wu, Xiangcong Yang, Yiyi Zhou, Chenxin Fang, Baiyang Song, Xiaoshuai Sun, and Rongrong Ji. Grounded chain-of-thought for multimodal large language models, 2025. URL https://arxiv.org/abs/2503.12799.

[71] Peng Xia, Ze Chen, Juanxi Tian, Yangrui Gong, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, Wenhao Zheng, Zhaoyang Wang, Xiao Wang, Xuchao Zhang, Chetan Bansal, Marc Niethammer, Junzhou Huang, Hongtu Zhu, Yun Li, Jimeng Sun, Zongyuan Ge, Gang Li, James Zou, and Huaxiu Yao. Cares: A comprehensive benchmark of trustworthiness in medical vision language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 140334–140365. Curran Associates, Inc., 2024. doi: 10.52202/079017-4455. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/fde7f40f8ced5735006810534dc66b33-Paper-Datasets_and_Benchmarks_Track.pdf.

[72] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/xuc15.html.

[73] Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, Yu Sun, Junao Shen, Chaojun Wang, Can Hu, Siwei Pan, Jie Tan, Tingyang Xu, Deli Zhao, Hao Zhang, and Yu Rong. Lingshu: Generalist foundation model for unified multimodal medical understanding and reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–18, 2026. doi: 10.1109/TPAMI.2026.3730310.

[74] Yousef Yeganeh, Azade Farshad, Ioannis Charisiadis, Marta Hasny, Martin Hartenberger, Björn Ommer, Nassir Navab, and Ehsan Adeli. Latent drifting in diffusion models for counterfactual medical image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7685–7695, June 2025.

[75] Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 20554–20565. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/ecb287ff763c169694f682af52c1f309-Paper.pdf.

[76] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=2msbbX3ydD.

[77] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 818–833, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10590-1.

[78] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Development of a large-scale medical visual question-answering dataset. *Communications Medicine*, 4(1):277, 2024.

[79] Lingting Zhu, Noel Codella, Dongdong Chen, Zhenchao Jin, Lu Yuan, and Lequan Yu. Generative enhancement for 3d medical images, 2024. URL https://arxiv.org/abs/2403.12852.

[80] Qingqing Zhu, Qiao Jin, Tejas S. Mathai, Yin Fang, Zhizheng Wang, Yifan Yang, Maame Sarfo-Gyamfi, Benjamin Hou, Ran Gu, Praveen T. S. Balamuralikrishna, Kenneth C. Wang, Ronald M. Summers, and Zhiyong Lu. Ct-bench: A benchmark for multimodal lesion understanding in computed tomography, 2026. URL https://arxiv.org/abs/2602.14879.

# A Limitations and Broader Impacts

## A.1 Limitations

While our work establishes a rigorous framework for evaluating attribution faithfulness in medical LVLM reasoning, several scope decisions define the boundaries of this study and suggest natural directions for future research.

First, our evaluation focuses on chest X-ray (CXR) imaging. We chose CXR because it is currently the only medical modality for which both large-scale expert spatial annotations and region-localized counterfactual editing models are publicly available, which are the two prerequisites for causally validated attribution evaluation. For other medical imaging modalities such as CT and MRI, domain-specific editing models with region-specific capabilities remain unavailable [3, 79, 74, 2, 80]. The construction recipe underlying our framework is modality-agnostic and can be extended to other modalities as analogous spatial annotation datasets and editing tools become available.

Second, our evaluation samples are based on a binary visual question answering reformulation: "Is there evidence of [attribute] in the image?" This approach enables straightforward assessment of model correctness, which is essential for the causal filtering procedure, and provides a clean, controlled testbed for attribution evaluation. Richer clinical tasks, such as full report generation or multi-step diagnostic reasoning using report-rich datasets like MIMIC-CXR [28], represent important and complementary research directions. These extensions would require new validation tools beyond binary correctness checking, and we view them as natural next steps building on the foundation established here.

Third, our framework evaluates attribution faithfulness only on samples that the model answers correctly, which is a principled methodological choice. Ground-truth attribution can only be reliably defined when the prediction itself is correct, since attribution targets for incorrect predictions are inherently ambiguous and require distinct validation methodologies. Attribution analysis on incorrect predictions is a valuable complementary problem that deserves dedicated attention in future work.

Fourth, MedGround-Bench is anchored to expert annotations. Each retained region is verified to be necessary for the prediction, but it is not an exhaustive account of the evidence the model may use, and cases whose relevant evidence is not captured by the annotations, such as global cues like the cardiothoracic ratio, may be excluded by the causal filter.

Finally, MedFocus currently uses anatomical regions as concepts, which do not capture finer clinical descriptors such as lesion extent. Its token-level attribution is computed in a single forward pass conditioned on the original output rather than by re-decoding the output under each intervention. Finer-grained concept vocabularies and regeneration-based attribution are natural extensions.

## A.2 Broader Impacts

This work aims to improve the trustworthiness of LVLMs in clinical settings by enabling more reliable evaluation of visual attribution methods. The positive societal impacts include supporting safer deployment of medical AI through better-grounded explanations, enabling clinicians to verify model reasoning before acting on AI outputs, facilitating error detection in high-stakes diagnostic scenarios, and providing the research community with tools to develop and benchmark more faithful attribution methods. By grounding explanations in clinically interpretable anatomical concepts, MedFocus further offers attributions that can be readily inspected and discussed by clinicians, supporting collaborative human-AI decision making.

We also note several considerations for responsible use. Attribution methods, including ours, should be understood as tools for identifying visual evidence that influences model predictions rather than as exhaustive explanations of internal model reasoning, and practitioners should not treat attribution outputs as a substitute for clinical judgment. Additionally, the source datasets used to construct our benchmark may carry distributional biases inherent to their collection sites and patient populations, which could affect how attribution faithfulness conclusions generalize across demographic groups and clinical contexts. We release our benchmark openly to enable the community to scrutinize, extend, and improve upon this evaluation framework, and we encourage future work to study attribution behavior across diverse demographic and clinical subgroups.

# B Details of MedGround-Bench Construction

The following subsections detail the construction of MedGround-Bench, covering data sources, the construction procedure, design rationale and validity of the causal filtering, and detailed benchmark statistics.

## B.1 Data Sources and Preprocessing

MedGround-Bench is constructed from three publicly available chest X-ray (CXR) datasets with spatial annotations, including ImaGenome [69], VinDR-CXR [48], and PadChest-GR [21]. ImaGenome and VinDR-CXR are sourced from PhysioNet [24]. PadChest-GR is sourced from Kaggle with permissions from the original data providers. Since ImaGenome is built upon MIMIC-CXR [28], we obtain the corresponding CXR images from MIMIC-CXR-JPG [29] for consistent JPEG processing.

During preprocessing, we resize all images to $224 \times 224$ pixels, consistent with prior work [69]. This ensures a fair comparison among attribution methods and with bounding boxes originally labeled by human experts under the same resolution. We retain only samples with annotations of abnormalities. For samples with multiple pieces of spatial evidence of the same attribute, we merge annotations into a bounding box list for that attribute. From ImaGenome, we select samples with attributes in "disease" or "anatomical finding" categories, yielding 1,405 visual questions. VinDR-CXR contributes all 2,108 samples. PadChest-GR provides 2,657 questions from the first 2,000 patients to maintain comparable scale. We additionally construct a small training set of 144 questions from the last 100 PadChest-GR patients for hyperparameter tuning of certain baseline methods, ensuring no overlap with evaluation samples.

## B.2 Construction Procedure

The three-step causal filtering procedure described in Section 3.2 is implemented with the following prompt templates and design choices.

Each annotated finding is reformulated as a binary visual question:

> *"Is there evidence of `[attribute]` in the image?"*

In direct mode, we append:

> *"Answer directly with yes or no without any explanation."*

In reasoning mode, we append:

> *"Think step by step and answer with yes or no."*

For the correctness filtering, we query six open-source LVLMs from different model families (Qwen2.5-VL-3B, Qwen2.5-VL-7B, Gemma3-4B, Gemma3-12B, MedGemma-4B, MedGemma1.5-4B) on each question and retain only correct predictions for subsequent filtering.

For foreground editing, we prompt RadEdit [53] with the bounding box annotation as the editing region and the text prompt

> *"No `[attribute]`"*.

We retain samples where the model flips its answer, indicating that the annotated region causally drives the prediction.

For background editing, we create the inverse mask of the bounding box and prompt RadEdit with the same text prompt. We additionally generate a variant with prompt

> *"No abnormality"*.

We retain only samples where predictions remain unchanged in both cases, confirming that answer changes in foreground editing are caused by the annotated region specifically.

After this three-step filtering process, we obtain a model-specific subset of the original questions with verified causal alignment between annotated bounding boxes and model predictions.

### B.3 Benchmark Statistics

The filtering procedure partitions the original samples into three groups for each dataset, model, and output mode:

- Incorrect: the model answers the original question incorrectly.
- Correct & Ungrounded: the model answers the original question correctly but fails at least one causal filtering condition.
- Correct & Grounded: the model answers correctly, flips under foreground editing, and remains unchanged under both background edits. These samples form MedGround-Bench.

Table 4 reports the percentage of samples in each group. Percentages are computed relative to the initial number of questions for each dataset: 1,405 for ImaGenome, 2,108 for VinDR-CXR, and 2,657 for PadChest-GR. The results show that many correct predictions are not causally grounded in the expert-annotated region. This confirms the need for causal filtering rather than relying on expert boxes alone as attribution ground truth.

Figure 6 in the main text provides an additional post-hoc analysis of these filtering stages. It shows that alignment between MedFocus attributions and expert annotations generally increases from incorrect samples to correct-but-ungrounded samples and then to correct-and-grounded samples. This pattern is consistent with the intended effect of the filter, but the benchmark construction itself does not depend on MedFocus or any other attribution method.

Table 4: Breakdown of sample distribution for each dataset and model across the three categories in both direct and reasoning modes. Percentages are computed relative to the initial number of questions for each dataset.

| Dataset | Model | Direct | | | Reasoning | | |
|---|---|---|---|---|---|---|---|
| | | Incorrect | Correct & Ungrounded | Correct & Grounded | Incorrect | Correct & Ungrounded | Correct & Grounded |
| ImaGenome | Qwen2.5-VL-3B | 49.47% | 43.06% | 7.47% | 40.00% | 56.01% | 3.99% |
| | Qwen2.5-VL-7B | 58.58% | 37.51% | 3.91% | 55.87% | 40.14% | 3.99% |
| | Gemma3-4B | 15.30% | 81.71% | 2.99% | 7.97% | 87.47% | 4.56% |
| | Gemma3-12B | 23.49% | 69.61% | 6.90% | 14.02% | 80.43% | 5.55% |
| | MedGemma-4B | 42.14% | 38.01% | 19.86% | 48.47% | 33.95% | 17.58% |
| | MedGemma1.5-4B | 33.67% | 50.25% | 16.09% | 39.36% | 40.36% | 20.28% |
| VinDR-CXR | Qwen2.5-VL-3B | 73.86% | 24.72% | 1.42% | 50.52% | 45.64% | 3.84% |
| | Qwen2.5-VL-7B | 77.32% | 20.59% | 2.09% | 65.42% | 32.40% | 2.18% |
| | Gemma3-4B | 41.75% | 53.18% | 5.08% | 13.43% | 79.41% | 7.16% |
| | Gemma3-12B | 47.01% | 49.19% | 3.80% | 32.78% | 64.18% | 3.04% |
| | MedGemma-4B | 54.32% | 41.27% | 4.41% | 55.93% | 39.47% | 4.60% |
| | MedGemma1.5-4B | 50.66% | 45.59% | 3.75% | 53.27% | 42.50% | 4.22% |
| PadChest-GR | Qwen2.5-VL-3B | 75.42% | 22.17% | 2.41% | 49.12% | 46.56% | 4.33% |
| | Qwen2.5-VL-7B | 82.57% | 15.81% | 1.62% | 73.01% | 25.25% | 1.73% |
| | Gemma3-4B | 52.24% | 45.01% | 2.75% | 22.36% | 72.86% | 4.78% |
| | Gemma3-12B | 35.49% | 60.56% | 3.95% | 33.50% | 62.40% | 4.10% |
| | MedGemma-4B | 43.66% | 49.19% | 7.15% | 48.14% | 45.54% | 6.32% |
| | MedGemma1.5-4B | 31.80% | 61.87% | 6.32% | 50.28% | 42.91% | 6.81% |

Table 5 reports the number of retained samples after all filtering steps. Pooling across models and datasets yields 1,880 samples in MedGround-Bench-Direct and 2,060 samples in MedGround-Bench-Reason. Per-model retention rates range from approximately 1.5% to 20%, reflecting the strictness of the three causal checks.

### B.4 Interpreting the Retained Samples

The causal filter does not attempt to reconstruct the model's full internal reasoning process. Instead, it identifies samples for which the expert annotation can be used as a reliable, model-specific attribution target. For each retained sample, removing the annotated finding changes the model's answer, while analogous edits outside the annotation leave the answer unchanged. Thus, the annotated region is not only clinically relevant, but also necessary for the model's prediction under the counterfactual intervention used in benchmark construction.

Table 5: Number of retained samples per dataset after all filtering steps. The final benchmark contains 1,880 direct-answer samples and 2,060 reasoning samples.

| Dataset | Qwen2.5 -VL-3B | Qwen2.5 -VL-7B | Gemma3 -4B | Gemma3 -12B | MedGemma -4B | MedGemma 1.5-4B | Total |
|---|---|---|---|---|---|---|---|
| **Direct** | | | | | | | |
| ImaGenome | 105 | 55 | 42 | 97 | 279 | 226 | 804 |
| VinDR-CXR | 30 | 44 | 107 | 80 | 93 | 79 | 433 |
| PadChest-GR | 64 | 43 | 73 | 105 | 190 | 168 | 643 |
| **Reasoning** | | | | | | | |
| ImaGenome | 56 | 56 | 64 | 78 | 247 | 285 | 786 |
| VinDR-CXR | 81 | 46 | 151 | 64 | 97 | 89 | 528 |
| PadChest-GR | 115 | 46 | 127 | 109 | 168 | 181 | 746 |

This definition still allows the model to rely on additional visual cues elsewhere in the image. The benchmark only requires that the annotated region be causally relevant, not that it be the sole source of evidence. Attribution methods are then evaluated on whether they can recover this verified evidence from the original image-question pair and model output. The expert boxes and RadEdit edits are used only to construct the benchmark, and are not provided to any evaluated attribution method. Methods that use image interventions define their own regions on the original image without access to RadEdit-inpainted expert boxes.

# C Implementation Details of Baseline Methods

Below, we provide detailed descriptions of the baseline methods used in our experiments. All gradient- and attention-based methods produce pixel-level saliency maps. For fair comparison with the bounding-box ground truth, we apply a standardized conversion. Each saliency map is first min-max normalized to $[0, 1]$ and thresholded at the 90th percentile of its non-zero values. We then extract 8-connected components from the resulting binary mask, discard components covering fewer than 16 pixels, and take the tight axis-aligned bounding box of each remaining component. The components are ranked by their mean saliency, and at most the top 10 bounding boxes per image are retained. The resulting boxes are rescaled to the native image resolution and evaluated against expert annotations via union-region overlap. This procedure is fixed across all methods and models. For the tuned-threshold comparison in Table 2, we recompute the baseline saliency maps and select the quantile threshold separately for each baseline and LVLM from $\{0.50, 0.60, 0.70, 0.75, 0.80, 0.85, 0.90, 0.925, 0.95, 0.975, 0.99\}$ by maximizing the mean IoU on the 144 held-out PadChest-GR tuning questions (Appendix B.1), keeping all other steps unchanged.

**Attention-based Methods.** We consider three purely attention-based attribution approaches: *(i)* **Attention Head**, which directly uses the attention weights from a selected head in the selected layer of the LVLM; *(ii)* **Attention Rollout** [1], which recursively multiplies attention matrices across layers to approximate token-level relevance; and *(iii)* **LRP (Layer-wise Relevance Propagation)** [6], which propagates attention-based relevance scores backward through the network using conservation rules. For all three methods, we select the best-performing layer or head using the training set of 144 questions from PadChest-GR described in Section B.1, evaluating based on average IoU across all samples for each model.

**Gradient-based Methods.** We compare with four gradient-based attribution techniques: *(i)* **Grad-CAM** [59], which produces class-discriminative localization maps via gradient-weighted activations of hidden states; *(ii)* **GradCAM++** [14], an improved variant that uses higher-order gradients for more accurate spatial attribution; *(iii)* **Gradient-weighted Attention** [15], which re-weights attention maps by the gradient signal; and *(iv)* **Integrated Gradients** [63], which accumulates gradients along a straight-line path from a baseline input to the actual input. For the first three methods, we select the best-performing layer using the same training set and evaluation procedure as described for attention-based methods. For Integrated Gradients, we select the best baseline strategy (zero image or mean pixel value) on the training set using the same evaluation procedure, with 36 integration steps.

**Prompting-based Methods.** We further compare with two prompting-based pipelines: *(i)* **Prompting**, which prompts the LVLM to directly output bounding box coordinates for the region most relevant to its prediction using the template:

> *"Identify the local evidence in the image that supports the answer, and output the bounding box coordinates. Provide your answer as a list of bounding boxes in the format* `[[x1, y1, x2, y2], ...]`*, where* `(x1, y1)` *is the top-left corner and* `(x2, y2)` *is the bottom-right corner of each bounding box."*

*(ii)* **Prompting + MedSAM** [44], which uses the VLM-identified region descriptions to prompt MedSAM for refined segmentation, employing the template:

> *"Identify the local evidence in the image that supports the answer, and output descriptions of the target objects/regions. Provide your answer as a list of words or phrases [*`"region1", "region2", ...`*] that concisely describe the target regions in the image."*

**Perturbation-based Methods.** We include two perturbation-based approaches: *(i)* **Occlusion** [77], which systematically slides a spatial patch over the input image and measures the change in model output to construct an importance map; and *(ii)* **RISE** [54], a method that estimates importance maps by probing the model with randomly masked versions of the input image. For both methods, we use $8 \times 8$ pixel patches as the unit of perturbation and replace masked patches with black pixels. RISE samples 64 random mask combinations per image, with 50% of patches masked in each combination. Unlike our concept-guided causal attribution, which performs structured interventions on semantically meaningful anatomical regions, both Occlusion and RISE apply spatially uniform perturbations without leveraging domain knowledge, treating all spatial locations as equally important units of perturbation.

# D Implementation Details of MedFocus

**Concept Vocabulary and Composite Groups.** Our method uses 11 predefined anatomical concepts from the ImaGenome dataset: cardiac silhouette, left lung, right lung, mediastinum, upper mediastinum, left clavicle, right clavicle, left hilar structures, right hilar structures, left costophrenic angle, and right costophrenic angle. These regions are routinely used by radiologists to interpret CXR images. We evaluate four clinically meaningful composite concept groups by masking the union of their bounding boxes: (1) left lung + right lung, (2) left clavicle + right clavicle, (3) left hilar structures + right hilar structures, and (4) left costophrenic angle + right costophrenic angle.

**Vocabulary granularity.** The attribution granularity of MedFocus is determined by the chosen concept vocabulary. In this work, we use ImaGenome anatomical regions because they provide a standardized and clinically interpretable concept set across CXR images. However, MedFocus is not restricted to these regions. For finer-grained settings, the vocabulary can be expanded to include concepts such as lung zones, lesion-level proposals, or measurement-related composite concepts, provided that reliable concept masks or proposals are available. Thus, limitations for findings such as small nodules, diffuse bilateral disease, or cardiothoracic-ratio-based cardiomegaly reflect the granularity of the current concept vocabulary rather than a structural constraint of the framework.

**Unbalanced Optimal Transport via Sinkhorn Algorithm.** We solve the unbalanced optimal transport problem using the Sinkhorn algorithm with entropic regularization [19, 10, 17]. Based on the original UOT objective in Equation (1), the regularized UOT problem is formulated as:

$$\mathbf{T}^* = \arg\min_{\mathbf{T}\geq 0} \sum_{i,j} C_{ij}\, T_{ij} + \varepsilon\, D_{\mathrm{KL}}(\mathbf{T} \| \mu_{\text{ref}} \otimes \mu_{\text{tgt}}) + \lambda_1\, D_{\mathrm{KL}}(\mathbf{T}\mathbf{1} \| \mu_{\text{ref}}) + \lambda_2\, D_{\mathrm{KL}}(\mathbf{T}^\top \mathbf{1} \| \mu_{\text{tgt}}), \tag{6}$$

where $\otimes$ is the outer product and the additional hyperparameter $\varepsilon > 0$ controls the smoothness of the transport plan. The Gibbs kernel $\mathbf{K} \in \mathbb{R}^{N \times M}$ has entries $K_{ij} = \exp(-C_{ij}/\varepsilon)$, where $N$ and $M$ are the number of pixels in the reference and target images. The Sinkhorn iterations alternate between

updating dual scaling variables $\mathbf{u} \in \mathbb{R}^N$ and $\mathbf{v} \in \mathbb{R}^M$:

$$\mathbf{u}^{(\ell+1)} = \left( \frac{\mu_{\text{ref}}}{\mathbf{K}\mathbf{v}^{(\ell)}} \right)^{\frac{\lambda_1}{\lambda_1+\varepsilon}}, \tag{7}$$

$$\mathbf{v}^{(\ell+1)} = \left( \frac{\mu_{\text{tgt}}}{\mathbf{K}^\top \mathbf{u}^{(\ell+1)}} \right)^{\frac{\lambda_2}{\lambda_2+\varepsilon}}, \tag{8}$$

with element-wise operations. Both scaling vectors are initialized as $\mathbf{u}^{(0)} = \mathbf{1}_N$ and $\mathbf{v}^{(0)} = \mathbf{1}_M$. After convergence at iteration $L$, the optimal transport plan is $T^*_{ij} = u_i^{(L)} \cdot K_{ij} \cdot v_j^{(L)}$. For each anatomical concept $c$ with reference pixel set $\mathcal{S}_c^{\text{ref}}$, we aggregate the transported mass at every target pixel, $m_j = \sum_{i \in \mathcal{S}_c^{\text{ref}}} T^*_{ij}$, and define the transferred region $\mathcal{S}_c^{\text{tgt}}$ as the smallest set of target pixels whose cumulative mass covers $75\%$ of the total: $\mathcal{S}_c^{\text{tgt}} = \{j_{(1)}, \ldots, j_{(k)}\}$ where $j_{(1)}, j_{(2)}, \ldots$ are the target pixels ranked by $m_j$ in descending order and $k$ is the smallest index satisfying $\sum_{r \le k} m_{j_{(r)}} \ge 0.75 \sum_j m_j$. This dense-core selection isolates the region that receives the bulk of mass from $\mathcal{S}_c^{\text{ref}}$ rather than the (numerically dense) full support of $T^*$ that the entropic regularizer produces.

**Reference Image Selection.** We select the reference normal CXR from ImaGenome by filtering all normal images with complete annotations covering all 11 concepts, yielding 16 candidates. For each candidate, we compute the UOT cost to the target image and select the one with the lowest total transport cost $\sum_{i,j} C_{ij} T^*_{ij}$. To reduce computational cost, both candidate and target images are downsampled to $14 \times 14$ resolution during this selection step. The full concept transfer is then performed at $56 \times 56$ resolution using the selected reference.

**Hyperparameters and Post-processing.** We set $\varepsilon = 0.05$ and $\lambda_1 = \lambda_2 = 0.1$ across all experiments. Sinkhorn iterations run for a maximum of $L = 500$ iterations or until the change in scaling vectors falls below $10^{-6}$. The UOT computation uses $56 \times 56$ downsampled images, with the final mapped concept masks upsampled to $224 \times 224$ for the MedSAM refinement step and subsequent causal attribution. The transferred concept mask for each concept is converted to a bounding box, which then serves as the prompt for MedSAM to produce a refined segmentation mask. After the causal attribution described in Section 4.2, the final attribution map is upsampled to the original image resolution of the CXR image, ensuring that the evaluation against expert-annotated bounding boxes is performed at the native resolution.

All experiments are conducted on NVIDIA A100 GPUs.

# E Further Quantitative Discussion

## E.1 Details of Model-specific Performance

Table 6 summarizes attribution performance of MedFocus across datasets and evaluation modes for each model. Larger models within the same family generally achieve higher attribution scores, indicating improved spatial grounding with increased model capacity. Medically trained models (MedGemma-4B and MedGemma1.5-4B) consistently outperform their non-medical counterparts (Gemma3-4B and Gemma3-12B), highlighting the benefit of domain-specific pretraining. Among MedGemma variants, the newer MedGemma1.5-4B achieves better results than the original MedGemma-4B, demonstrating the impact of continued model improvements. These trends are observed across both direct and reasoning modes and are consistent across all datasets.

## E.2 Comparison of Method Efficiency

Table 7 compares the average inference time per sample for all attribution methods. Attention-based approaches are the fastest overall, followed closely by GradCAM, GradCAM++, and Gradient-weighted Attention. Prompting-based methods take approximately one second per sample. MedFocus requires 1.65 seconds per sample, making it slower than lightweight gradient- and attention-based baselines but still substantially faster than the more expensive perturbation-based alternatives, including Occlusion, RISE, and especially Integrated Gradients. Although MedFocus is not the cheapest method computationally, it offers a strong efficiency-faithfulness trade-off by combining clearly superior attribution quality with a runtime that remains practical.

Table 6: Model-specific attribution performance across datasets and evaluation modes.

| Model | ImaGenome | | | | VinDR-CXR | | | | PadChest-GR | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | IoU | F1 | Prec | Recall | IoU | F1 | Prec | Recall | IoU | F1 | Prec | Recall |
| **Direct Mode** | | | | | | | | | | | | |
| Qwen2.5-VL-3B | 49.60 | 63.65 | 70.62 | 67.67 | 21.52 | 32.94 | 24.81 | 72.31 | 32.76 | 45.54 | 45.38 | 69.07 |
| Qwen2.5-VL-7B | 48.10 | 61.10 | 65.07 | 68.22 | 18.48 | 28.15 | 20.14 | 85.23 | 38.48 | 51.72 | 55.75 | 61.41 |
| Gemma3-4B | 39.33 | 52.66 | 54.93 | 64.43 | 10.65 | 17.21 | 11.97 | 65.87 | 22.37 | 33.30 | 33.41 | 48.93 |
| Gemma3-12B | 43.96 | 57.89 | 57.99 | 69.60 | 12.70 | 19.98 | 13.03 | 82.90 | 29.58 | 42.40 | 38.41 | 72.44 |
| MedGemma-4B | 58.16 | 71.13 | 65.65 | 84.94 | 16.21 | 24.72 | 17.08 | 85.13 | 33.38 | 46.02 | 39.09 | 77.12 |
| MedGemma1.5-4B | 60.38 | 73.32 | 65.21 | 90.05 | 16.63 | 25.76 | 17.55 | 89.43 | 38.08 | 51.28 | 42.83 | 84.58 |
| **Reasoning Mode** | | | | | | | | | | | | |
| Qwen2.5-VL-3B | 40.73 | 54.26 | 48.94 | 77.41 | 13.01 | 21.16 | 13.97 | 85.02 | 23.16 | 34.27 | 26.10 | 71.36 |
| Qwen2.5-VL-7B | 46.74 | 60.39 | 60.07 | 71.20 | 11.80 | 19.05 | 12.21 | 78.18 | 30.37 | 41.24 | 37.94 | 62.31 |
| Gemma3-4B | 40.61 | 53.94 | 45.51 | 84.48 | 4.89 | 8.80 | 5.03 | 75.43 | 19.41 | 29.03 | 22.07 | 75.97 |
| Gemma3-12B | 43.58 | 56.72 | 49.65 | 81.49 | 6.39 | 10.29 | 6.52 | 80.83 | 21.90 | 32.48 | 24.24 | 82.61 |
| MedGemma-4B | 58.04 | 71.21 | 63.14 | 90.10 | 14.26 | 21.57 | 14.56 | 94.72 | 32.18 | 44.08 | 35.82 | 82.88 |
| MedGemma1.5-4B | 57.49 | 70.90 | 63.29 | 88.24 | 15.18 | 23.71 | 15.49 | 95.30 | 34.62 | 47.63 | 38.17 | 86.41 |

Table 7: Inference time (seconds per sample) for each visual attribution method, grouped by method category.

| | **Gradient-based** | | | | **Prompting-based** | |
|---|---|---|---|---|---|---|
| | GradCAM | GradCAM++ | Integrated Gradients | Gradient-weighted Attn. | Prompting | Prompting + MedSAM |
| Time | 0.53 | 0.61 | 7.60 | 0.60 | 1.09 | 0.98 |
| | **Attention-based** | | | **Perturbation-based** | | |
| | Attention Head | Attention Rollout | LRP | Occlusion | RISE | MedFocus (Ours) |
| Time | 0.42 | 0.43 | 0.40 | 2.64 | 2.49 | 1.65 |

### E.3 Hyperparameter Sensitivity Analysis

Table 8 presents a joint sensitivity analysis of MedFocus across key UOT hyperparameters: the number of candidate reference images (#C), the marginal relaxation coefficients ($\lambda_1, \lambda_2$), and the entropic regularization coefficient ($\varepsilon$). The results demonstrate that MedFocus exhibits reasonable stability across a range of hyperparameter settings. Specifically, increasing $\lambda$ (e.g., $\lambda = 1.0$) enforces stricter adherence to the original marginal distributions, yielding degraded performance compared to more relaxed settings (e.g., $\lambda = 0.1$ or $\lambda = 0.01$). Similarly, larger $\varepsilon$ values (e.g., $\varepsilon = 0.5$) produce overly smooth transport plans with high recall but low precision, whereas smaller values (e.g., $\varepsilon = 0.005$) generate sharper plans with higher precision at the cost of lower recall. Our selection of $\varepsilon = 0.05$ and $\lambda = 0.1$ achieves a well-balanced trade-off between precision and recall. Notably, performance remains stable as the number of candidate reference images decreases, indicating that MedFocus is robust to baseline selection and does not require a large candidate pool to achieve strong results.

### E.4 Concept Frequency Analysis

Table 9 reports the frequency with which MedFocus identifies each anatomical concept as the most important evidence source across MedGround-Bench-Direct and MedGround-Bench-Reason. The left and right lungs dominate, accounting for the vast majority of attributions across all three datasets. This pattern is expected, as most benchmark questions concern pulmonary findings localized within the lung fields. The cardiac silhouette appears more frequently on PadChest-GR than on ImaGenome or VinDR-CXR, reflecting the higher prevalence of cardiac and mediastinal findings in PadChest-GR. In contrast, smaller or more peripheral concepts, such as the clavicles, costophrenic angles, and upper mediastinum, are rarely selected. A notable observation is that the reasoning setting shows an even stronger concentration on lung concepts than the direct setting, particularly for ImaGenome and PadChest-GR. This suggests that when models generate intermediate rationales, they tend to attend

Table 8: Joint sensitivity analysis of UOT hyperparameters: number of candidate reference images (#C), marginal relaxation $\lambda$ ($= \lambda_1 = \lambda_2$), and entropic regularization $\varepsilon$. IoU, F1, Precision, and Recall are averaged across all datasets and models.

| $\lambda$ | $\varepsilon$ | #C = 1 | | | | #C = 4 | | | | #C = 16 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **IoU** | **F1** | **Prec** | **Recall** | **IoU** | **F1** | **Prec** | **Recall** | **IoU** | **F1** | **Prec** | **Recall** |
| **0.01** | 0.005 | 35.71 | 47.24 | 50.32 | 64.52 | 35.00 | 46.54 | 51.57 | 61.68 | 35.00 | 46.55 | 51.54 | 61.74 |
| | 0.05 | **38.17** | **49.84** | 46.48 | 75.56 | **37.80** | **49.59** | 47.02 | 72.69 | 37.78 | 49.58 | 47.01 | 72.81 |
| | 0.5 | 35.77 | 47.94 | 39.00 | 87.69 | 36.30 | 48.43 | 39.70 | 87.95 | 36.38 | 48.50 | 39.77 | 87.98 |
| **0.1** | 0.005 | 35.79 | 47.14 | 48.41 | 66.88 | 35.96 | 47.36 | 51.21 | 63.89 | 35.93 | 47.34 | 51.28 | 63.74 |
| | 0.05 | 36.80 | 48.52 | 44.68 | 74.98 | 37.38 | 49.33 | 45.88 | 75.06 | **37.93** | **49.82** | 45.37 | 77.68 |
| | 0.5 | 35.20 | 47.47 | 38.30 | 88.42 | 35.90 | 48.06 | 38.82 | 89.65 | 35.88 | 48.04 | 38.81 | 89.57 |
| **1.0** | 0.005 | 34.45 | 45.65 | 46.72 | 63.78 | 34.78 | 46.00 | 49.86 | 61.00 | 34.80 | 46.04 | 49.78 | 61.01 |
| | 0.05 | 34.61 | 46.55 | 43.17 | 71.51 | 34.93 | 47.00 | 44.15 | 71.31 | 34.90 | 46.96 | 44.09 | 71.42 |
| | 0.5 | 34.23 | 46.38 | 36.28 | 91.59 | 34.40 | 46.48 | 36.27 | 92.25 | 34.41 | 46.51 | 36.29 | 92.31 |

to broader regions in the CXR to support their reasoning. Overall, the concept-frequency analysis provides an interpretable characterization of where LVLMs ground their medical predictions and identifies the anatomical regions most influential to benchmark performance.

Table 9: Frequency of anatomical concepts identified by MedFocus as important for LVLM outputs across MedGround-Bench-Direct and MedGround-Bench-Reason.

| | MedGround-Bench-Direct | | | MedGround-Bench-Reason | | |
|---|---|---|---|---|---|---|
| **Concept** | **ImaGenome** | **VinDR-CXR** | **PadChest-GR** | **ImaGenome** | **VinDR-CXR** | **PadChest-GR** |
| Cardiac silhouette | 3.48% | 4.62% | 8.09% | 1.65% | 4.17% | 5.50% |
| Left lung | 75.12% | 59.35% | 60.03% | 87.15% | 79.55% | 80.70% |
| Right lung | 73.51% | 53.81% | 57.70% | 87.40% | 76.89% | 78.28% |
| Mediastinum | 7.96% | 10.39% | 9.64% | 2.42% | 5.87% | 6.03% |
| Upper mediastinum | 1.00% | 2.54% | 2.18% | 0.25% | 1.70% | 0.67% |
| Left clavicle | 2.11% | 6.93% | 4.35% | 1.91% | 1.14% | 0.80% |
| Right clavicle | 1.99% | 5.54% | 4.04% | 2.04% | 0.76% | 0.67% |
| Left hilar structures | 4.35% | 8.78% | 8.09% | 2.04% | 3.60% | 2.01% |
| Right hilar structures | 5.10% | 7.85% | 7.31% | 2.67% | 3.60% | 2.41% |
| Left costophrenic angle | 2.11% | 2.77% | 2.64% | 0.38% | 0.19% | 0.67% |
| Right costophrenic angle | 1.49% | 2.77% | 2.64% | 0.25% | 0.00% | 0.67% |

### E.5 Further Discussion on Intervention Strategies

Beyond the comparison with healthy-tissue inpainting by RadEdit in Section 5.4 ("Bounding box + RadEdit" in Table 3), we further discuss cross-disease replacement as an alternative intervention, in which RadEdit is prompted with a different finding in each concept region (e.g., pneumothorax for a question about atelectasis). Because generating RadEdit edits for every concept is computationally expensive, this comparison is conducted on a subset of MedGround-Bench-Direct. For a fair comparison, all interventions are evaluated on samples from the same subset. As shown in Table 10, zero masking achieves the best IoU, F1, and recall.

Table 10: MedFocus attribution performance under different intervention types on a subset of MedGround-Bench-Direct. Best results are in **bold**.

| **Intervention** | **IoU** | **F1** | **Prec** | **Recall** |
|---|---|---|---|---|
| **Zero masking (default)** | **45.55** | **58.89** | 60.24 | **71.64** |
| Healthy-tissue inpainting | 41.08 | 54.44 | **61.45** | 62.75 |
| Cross-disease replacement | 32.16 | 45.02 | 57.92 | 47.44 |

To test whether the advantage of zero masking stems from generic image corruption, we add a matched control that zero-masks a box of the same size as the concept selected by MedFocus, placed at a non-overlapping location obtained by reflecting the box through the image center. As shown in Table 11, masking the selected concept reduces the probability of the original answer much more than masking the control region, indicating that the model responds to the removal of concept-specific evidence rather than to the masking artifact itself.

Table 11: Probability of the original answer after zero masking the concept selected by MedFocus or a same-size control region that does not overlap it, on the same subset as Table 10.

| Zero-masked region | Mean | Median |
|---|---|---|
| Concept selected by MedFocus | 0.22 | 0.09 |
| Same-size control region | 0.57 | 0.70 |

### E.6 Evaluation on an Additional Medical LVLM

To broaden the coverage of medical LVLMs, we apply the MedGround-Bench construction pipeline to Lingshu-7B [73], a medical LVLM fine-tuned from Qwen2.5-VL-7B, in the direct-answer mode. This retains 303 samples (172 from ImaGenome, 41 from VinDR-CXR, and 90 from PadChest-GR), which are evaluated separately and not added to MedGround-Bench. Table 12 shows that Lingshu-7B answers more questions correctly and yields about twice the share of correct and grounded samples of its base model.

Table 12: Distribution of filtering outcomes for Lingshu-7B and its base model Qwen2.5-VL-7B in the direct-answer mode, combined across the three datasets.

| Model | Incorrect (%) | Correct & ungrounded (%) | Correct & grounded (%) |
|---|---|---|---|
| Qwen2.5-VL-7B | 75.3 | 22.4 | 2.3 |
| Lingshu-7B | 66.3 | 28.8 | 4.9 |

Table 13 compares MedFocus with representative attention-, gradient-, and perturbation-based baselines on these samples. MedFocus achieves the best IoU and F1 on all three datasets, consistent with the results on the six original LVLMs.

Table 13: Attribution performance on the 303 direct-answer samples retained for Lingshu-7B. All scores are percentages. Best results are in **bold**.

| Method | ImaGenome | | | | VinDR-CXR | | | | PadChest-GR | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | IoU | F1 | Prec | Recall | IoU | F1 | Prec | Recall | IoU | F1 | Prec | Recall |
| Attention Head | 9.96 | 17.43 | 37.42 | 11.73 | 8.27 | 14.51 | 16.38 | 21.95 | 10.15 | 16.60 | 27.88 | 14.40 |
| GradCAM [59] | 8.87 | 15.39 | 42.19 | 11.04 | 13.31 | 21.55 | **37.09** | 21.49 | 13.15 | 21.13 | 38.95 | 17.44 |
| Occlusion [77] | 23.67 | 37.03 | 70.17 | 26.78 | 20.12 | 31.58 | 35.52 | 47.45 | 25.02 | 37.78 | **61.38** | 33.06 |
| **Ours (MedFocus)** | **64.83** | **77.23** | **70.49** | **90.60** | **26.66** | **38.36** | 29.75 | **91.64** | **37.34** | **50.67** | 50.06 | **70.83** |

### E.7 Multi-region Evidence and Per-region Performance

**Multi-region evidence.** Because the counterfactual edits in MedGround-Bench act on the union of all annotated regions, cases relying on several regions are retained whenever the combined edit changes the prediction and the background edit does not. Table 14 compares the number of annotated regions, counted over the 11 anatomical concepts, in the ImaGenome source questions and in the retained samples. Multi-region cases are not underrepresented after filtering: questions spanning more than two regions account for 27.8% of the source questions but 44.7% of the retained samples.

Table 14: Distribution of the number of annotated regions on ImaGenome before and after filtering.

| # Annotated regions | Source (%) | Retained (%) |
|---|---|---|
| 1 | 37.1 | 20.3 |
| 2 | 35.1 | 35.1 |
| $>2$ | 27.8 | 44.7 |

**Bilateral evidence.** When a prediction relies on both lungs, MedFocus can select the composite left lung + right lung concept. We measure how well each method recovers such bilateral evidence on ImaGenome in the direct-answer mode. An attribution is counted as bilateral if it covers at least 25% of each lung field. Recall is the fraction of samples with bilateral annotations whose attribution is

bilateral, and precision is the fraction of samples with bilateral attributions whose annotations are also bilateral. As shown in Table 15, MedFocus recovers bilateral evidence far more often than the baselines, whose thresholded boxes rarely cover both lungs.

Table 15: Identification of bilateral (both-lung) evidence on ImaGenome in the direct-answer mode. Best results are in **bold**.

| Method | Recall | Precision | F1 |
|---|---|---|---|
| Attention Head | 5.7 | **82.1** | 10.7 |
| GradCAM [59] | 57.6 | 69.6 | 63.1 |
| Occlusion [77] | 37.0 | 77.4 | 50.1 |
| **Ours (MedFocus)** | **87.8** | 74.0 | **80.3** |

**Per-region performance.** Table 16 reports IoU on ImaGenome in the direct-answer mode, grouped by the anatomical regions covered by the annotation. A sample whose annotation spans several regions is counted in each of them. MedFocus achieves the highest IoU in every region.

Table 16: IoU by anatomical region on ImaGenome in the direct-answer mode. The number of samples whose annotation covers each region is given in parentheses. Best results are in **bold**.

| Method | Lungs (767) | Hilar structures (289) | Costophrenic angles (205) | Cardiac silhouette (35) |
|---|---|---|---|---|
| Attention Head | 16.14 | 17.32 | 16.87 | 20.27 |
| GradCAM [59] | 29.04 | 30.18 | 29.65 | 22.74 |
| Occlusion [77] | 24.00 | 25.96 | 26.96 | 39.10 |
| **Ours (MedFocus)** | **55.32** | **61.56** | **56.32** | **45.39** |

# F Qualitative Model Comparison and Error Analysis

Figure 7 compares MedFocus attributions across Gemma3 and MedGemma variants on three representative examples from each source dataset: aspiration from ImaGenome, nodule/mass from VinDR-CXR, and abnormal foreign body or metal from PadChest-GR. The examples reveal a clear qualitative gap between general-purpose and medically trained models across the examples. On aspiration, the MedGemma variants localize the abnormal bilateral lung regions more precisely and achieve substantially higher overlap with expert annotations than the Gemma3 variants. On nodule/mass, however, all models produce overly broad lung-level attributions rather than tightly focusing on the small focal lesion. This pattern is even more pronounced for abnormal foreign body or metal, where the true evidence occupies a very small spatial region and all models partially collapse to coarse thoracic-level attributions. These cases reveal two recurring error patterns in MedFocus attributions derived from the underlying LVLMs. First, attributions can be overly broad, correctly identifying the general anatomical region while failing to localize the lesion precisely. Second, attributions can exhibit partial coverage, overlapping the annotated region but missing part of the supporting evidence. Taken together with the results in Figures 4 and 7, these examples show that MedFocus substantially outperforms existing baselines, while still leaving room for improvement in achieving better coverage for weaker models and sharper localization for small or highly focal findings.

Figure 8 provides a finer-grained comparison in the reasoning setting using an example about osteosynthesis material from PadChest-GR. Although all four models answer the question correctly, the quality of their reasoning-grounding alignment differs substantially. The Gemma3 models rely on partially relevant but diffuse evidence and include more generic descriptions of the image. In contrast, the MedGemma models concentrate more directly on the left shoulder/clavicular region containing the hardware and produce more clinically specific reasoning, referring to metallic density and hardware-like structures. The token-level concept attribution further shows that the words most affected by intervention are visually grounded near the annotated region for the stronger models, whereas weaker models distribute importance across broader, less specific areas. Together, Figures 7 and 8 show that correct answering alone is insufficient. Models differ markedly in how well their spatial attributions and intermediate reasoning are tied to the true supporting evidence.

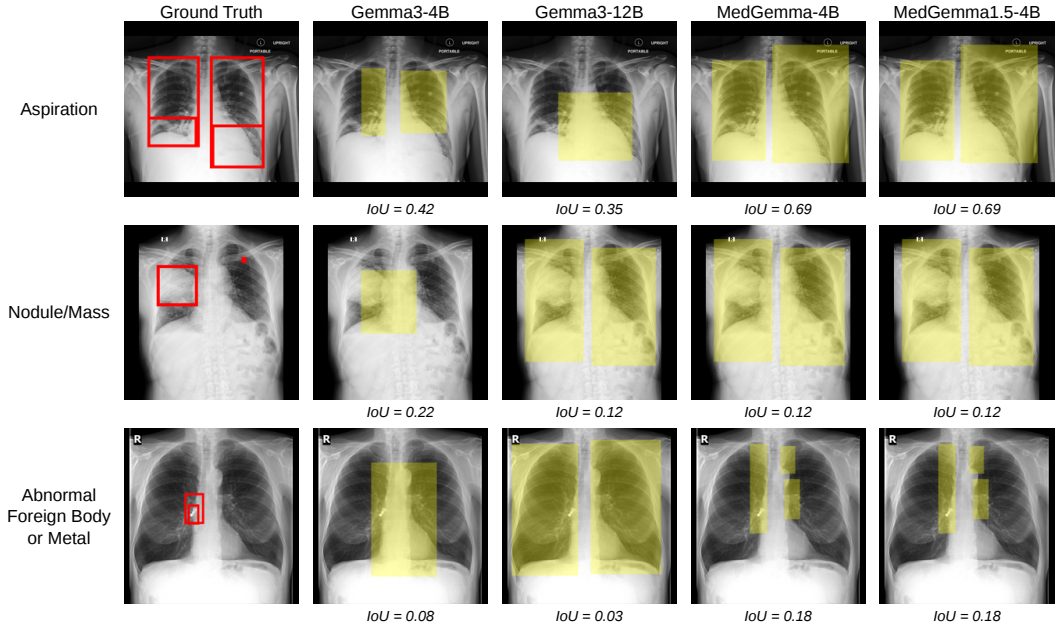


Figure 7: Qualitative comparison of MedFocus spatial attributions across Gemma3 and MedGemma variants on three representative MedGround-Bench examples. Ground-truth evidence is shown in red and predicted attributions are shown in yellow.

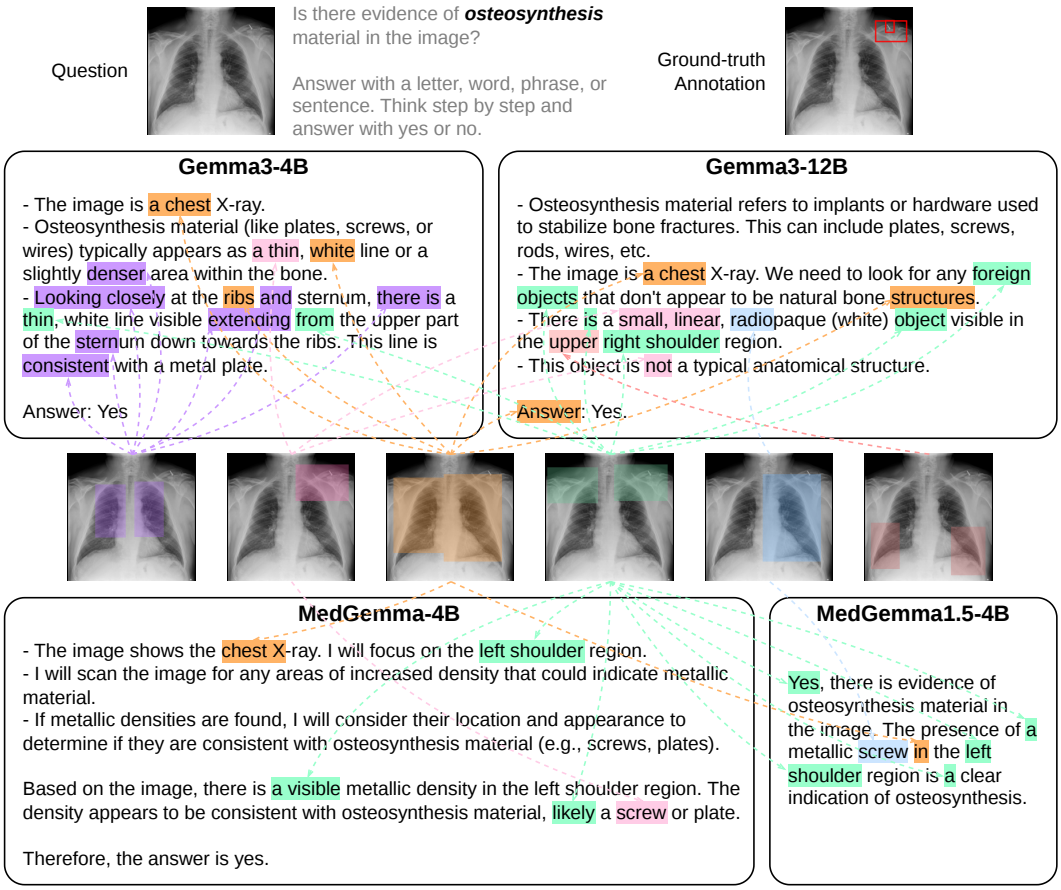


Figure 8: Token-level concept attribution for a reasoning example about osteosynthesis material. Colored words indicate tokens whose probabilities are most affected by concept intervention, and the corresponding highlighted regions show the attributed evidence.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Appendix A.1.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: See Appendices B, C, and D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code and curated benchmark are released at the URL provided in the abstract.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: See Appendices C and D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We aggregate metrics across six models and three datasets, which provides robustness across model and data variation.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Appendices D and E.2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See Appendix A.2.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We release a benchmark constructed from existing public datasets and an attribution method. we do not release new generative or diagnostic models.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The creators or original owners of assets used in the paper are all properly credited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: See Appendix B and the URL provided in the abstract.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: There is no new human subjects research in this paper.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: There is no new human subjects research in this paper.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLM is not an important, original, or non-standard component of the core methods in this research.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.