Attention Without Grounding: Causal Evaluation of Visual Explanations in Medical VLMs

Binesh Sadanandan and Vahid Behzadan

SAIL Lab, University of New Haven, West Haven, CT, USA
{bsada1@unh,vbehzadan@}newhaven.edu

Abstract. Attention and saliency heatmaps are increasingly shown as explanations of medical Vision-Language Model (VLM) outputs on chest X-rays (CXR), yet whether a heatmap points to the image evidence that actually drives the answer has not been tested causally. We audit explanation faithfulness along three axes: overlap with radiologist bounding boxes on PadChest [4] (n=637), attribution mass inside radiologist pixel masks on CheXlocalize [12] (n=643), and 16×16 patch-occlusion causal maps that record which regions, when hidden, change the prediction. We evaluate three MedGemma-4B variants [21,18] with cross-family probes on LLaVA-RAD [1], Qwen3-VL-8B-Instruct [24], and the specialist CheXagent-2-3b [29]; two CXR-trained classifiers (DenseNet121, ResNet50) [22,23] are positive controls. **A faithful heatmap needs two things at once: the model must use the image, and its attention must fall on the regions whose occlusion changes the answer. No evaluated VLM satisfies both.** The MedGemma variants and the Qwen3-VL probe use the image, but their attention anti-correlates with patch-occlusion causal importance (−0.098 base, −0.050 targeted LoRA, −0.168 full LoRA, −0.117 Qwen3-VL; all 95% bootstrap confidence intervals below zero). LLaVA-RAD's attention correlates positively, but only because it barely uses the image (99.1% text-only agreement, near-zero causal mass), so its ρ joins two near-zero signals. Attention also misses the annotated anatomy: true-region overlap never beats shifted or random controls, and no method places more than 22% of its mass inside the radiologist mask. Two CXR-trained classifiers pass the same protocol ($p<10^{-4}$ over shuffled on every metric), so the failure is specific to the heatmaps, not the metric. These heatmaps are reassuring but not faithful: clinical explanations need controlled localization metrics and causal perturbation, not visual inspection alone.

Keywords: Visual explanation · Explanation faithfulness · Attention grounding · Causal perturbation · Medical VLM.

1 Introduction

When a radiology VLM says "cardiomegaly present," the natural follow-up is: what part of the image drove that answer? The usual reply is an attention

heatmap. If the hot region lands on the heart silhouette, the explanation feels reassuring. But reassurance is not evidence. We ask whether these heatmaps point to clinically relevant image evidence, and what causal probe can corroborate that judgment, an interpretability question that visual inspection alone cannot settle.

The faithfulness of attention as explanation is contested in language models [8,9] and limited for chest-X-ray classifiers [12]. Whether VLMs, which jointly attend over text and image tokens with generative decoding, produce more faithful explanations is unresolved; we extend the evaluation to multiple attribution methods, controlled bounding-box and pixel-mask baselines, and patch-level causal perturbation.

We study MedGemma-4B [21] in three configurations: the base model, a targeted Low-Rank Adaptation (LoRA) on layers 15–19 [18], and a full LoRA across all 34 layers. Three cross-family probes test whether the behaviour is architecture-specific. LLaVA-RAD [1] is a 7B Vicuna model with a BiomedCLIP-CXR encoder; Qwen3-VL-8B-Instruct [24] is a general-purpose VLM; and CheXagent-2-3b [29] is a chest-X-ray specialist. Two CXR-trained classifiers, DenseNet121-Chex and ResNet50-All, provide positive localization controls through Grad-CAM [22,23]. These families span a wide range of vision encoders and decoders, so a consistent failure cannot be pinned on one architecture. We evaluate on MIMIC-CXR [5], on PadChest [4] with radiologist bounding boxes for n=637 cases, and on CheXlocalize [12] with pixel masks across ten pathologies for n=643 cases.

Contributions. We make three claims, each established by causal perturbation rather than by visual inspection. *(i)* In every evaluated medical VLM, attention and saliency heatmaps do not track causal importance: true-region overlap does not exceed random controls, and the attention-causal Spearman ρ is negative for the MedGemma variants and Qwen3-VL, with LLaVA-RAD reversing sign only because its causal mass is near zero. *(ii)* The failure is not an artifact of method, modality, or metric: it holds across six attribution methods, including the axiomatic integrated gradients [27] and the model-agnostic RISE [25], while the same protocol detects strong localization in the CXR-trained classifier baselines. *(iii)* Grounding is achievable but not by these general medical VLMs: a CXR specialist is causally grounded on the annotation style it matches yet not on a second dataset, and fine-tuning that reduces paraphrase flips does not improve grounding.

What we do not claim. We do not claim that all VLM explanations fail, that the classifier baselines are clinically superior, or that we have exhausted the attribution-method space. The practical claim is narrow and falsifiable: visual inspection of these heatmaps is not a reliable test of grounding.

2 Related Work

Attention faithfulness is contested in NLP [8,9] and benchmarked for chest-X-ray classifiers [12,26,15]. Medical-VLM trustworthiness benchmarks measure

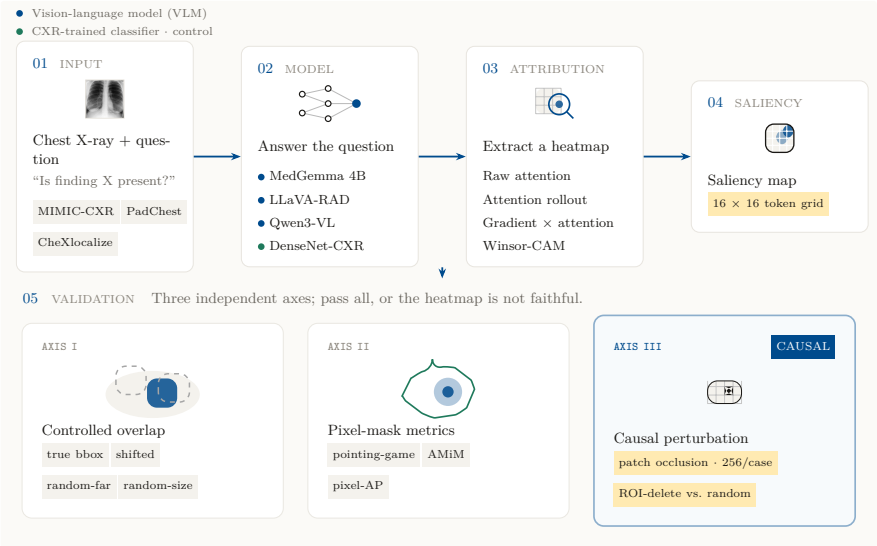

Fig. 1. Explanation-validation protocol. A saliency map (pipeline, top) is tested along three independent axes: (i) controlled overlap against shifted, random-far, and random-size bounding boxes on PadChest; (ii) pixel-mask metrics (pointing-game, attribution mass in mask, pixel-AP) against radiologist segmentations on CheXlocalize; (iii) causal perturbation via 16×16 patch occlusion (256 forward passes per case). A heatmap is faithful only if overlap exceeds the controls *and* high-saliency patches are causally important; in this audit every evaluated VLM fails at least one axis, while the CXR-trained classifier controls pass all three.

paraphrase sensitivity [17,3], modality bias [14], hallucination [11,10], and uncertainty under shift [19] individually. Concurrent work finds that vision-language benchmarks barely test visual grounding [30], that medical VLMs trade grounding for sycophantic agreement under social pressure [31], and that they default to anatomical priors on rare findings [32]; we add a direct causal-faithfulness audit of the heatmaps these models produce. Mechanistic work [18] traces paraphrase consistency to a layer-17 sparse autoencoder feature and reduces flips via targeted LoRA; we reuse that checkpoint and contribute the grounding, multi-method attribution, and counterfactual analyses.

3 Methods

3.1 Models and Datasets

We evaluate eight configurations across six model families (Table 1): MedGemma-4B-IT [21] in three configurations, two cross-family VLM probes, a chest-X-ray specialist VLM, and two CXR-trained classifier baselines.

Table 1. Models evaluated. LoRA adapters use rank 16, α=32, dropout 0.05, lr 2×10^{-4}, batch 8, at n=500 binary-presence and n=12K multi-task scales; the 98 presence test items are held out from both. Classifiers use class-specific Grad-CAM [22] upsampled to 224×224.

Model	Backbone / configuration	Role
MedGemma-4B Base	Gemma 2 + SigLIP; 256 tokens (16×16)	VLM
Targeted LoRA	rank-16 Q/K/V, L15–19; SAE feature 3818 @ L17 [6]	tuned
Full LoRA	rank-16 all projections, L0–33	tuned
LLaVA-RAD-7B [1]	Vicuna-7B + BiomedCLIP-CXR	probe
Qwen3-VL-8B-Instruct [24]	general-purpose dense VLM (8×8 grid)	probe
CheXagent-2-3b [29]	chest-X-ray specialist VLM	decision
DenseNet121-Chex [23]	`densenet121-res224-chex` @ `features.norm5`	control
ResNet50-All [23]	`resnet50-res512-all` (multi-dataset) @ `model.layer4`	control

Datasets. MIMIC-CXR [5] (98 binary presence questions, 3–5 paraphrases each, credentialed); PadChest [4] (861 binary questions, 637 with radiologist bounding boxes across eight finding groups, public); CheXlocalize [12] validation split (643 samples with pixel masks across ten pathologies, public); and VinDr-CXR [28] (n=300 presence questions with radiologist boxes) as a secondary replication set. Paraphrases come from template-based rules over five linguistic phenomena, filtered with BioClinicalBERT [20] cosine similarity > 0.95; no LLM-generated text is used.

Reproducibility. MedGemma-4B-IT [21], LLaVA-RAD [1], Qwen3-VL-8B-Instruct [24], and the DenseNet-CXR weights via TorchXRayVision [23] are available under their licenses; LoRA follows Hu et al. [2] defaults with layer selection from [18]. All metrics report 95% bootstrap CIs (2000 resamples). Per-pathology paired permutation tests (true vs. shuffled saliency, 10,000 resamples) are Holm-Bonferroni corrected across the ten CheXlocalize pathologies at α=0.05; pooled paired Cohen's d quantifies effect size. Even the strongest VLM configuration (full LoRA, gradient-attention) is heterogeneous: 8 of 10 pathologies are Holm-significant on at least one metric, the only exceptions being the two lowest-prevalence classes (Pneumothorax n=8, Lung Lesion n=1); the well-powered classes (n=33–125) drive the aggregate. Code and the evaluation protocol are at https://github.com/thedatasense/medicalvlm_attention_without_grounding; the n=12K LoRA adapters are a Hugging Face collection.[1] CheXagent-2-3b requires a separate Transformers 4.40 / float32 environment.

3.2 Evaluation Protocol

The protocol (Fig. 1) has three axes. **Grounding overlap:** attention from the last text token to the 256 image tokens, averaged over heads and layers 10–20 and thresholded at the 90th percentile, scored against the bounding-box mask versus true, shifted, random-far, and random-size controls (with a nine-group layer sweep). **Pixel-mask attribution:** four methods (raw attention, attention rollout [13], gradient-attention product [7], Winsor-CAM), upsampled to

[1] https://huggingface.co/collections/saillab/mechanistically-guided-lora-chil-2026-69ff7afccce7547a00180b2a

224×224 and scored against radiologist masks by pointing-game hit rate, attribution mass in mask (AMiM), and pixel-AP with paired permutation. **Counterfactual perturbation:** patch occlusion mean-fills each of the 256 patches and records the margin shift $\delta_{r,c} = m_{\text{orig}} - m_{\text{occluded}}$ (positive when occluding a patch reduces the answer margin) to build a 16×16 causal map; the main table correlates against the signed shift, the per-method comparison against $|\delta_{r,c}|$. We add integrated gradients [27] and RISE [25] on this axis (six methods total) and report insertion/deletion AUC. Mid-gray fill is used throughout, reported as a known protocol choice.

4 Experiments and Results

Behavioural context. Companion work [17,18] shows that reduced paraphrase sensitivity coincides with reduced image use (LLaVA-RAD: 99.1% text-only agreement; Full LoRA: up to 77.9%), motivating whether these heatmaps are grounded.

4.1 Attention Grounding Failure

Bounding-box overlap. For the base MedGemma model on PadChest (n=637), true-box attention coverage (0.037, 95% CI 0.031–0.044) is the *lowest* of four conditions, below shifted (0.042), random far (0.052), and random sized (0.048) boxes. The pattern holds for both LoRA variants and across a nine-group layer sweep, so it is not a layer-selection artifact. LLaVA-RAD is weaker still (true-box coverage 1.2%).

Multi-method evaluation on pixel masks (Table 2; Fig. 2). On CheXlocalize [12] pixel masks (n=643), no VLM attribution method places more than 22% of saliency mass inside the radiologist mask. Gradient-attention product reaches the highest VLM pointing-game (44–51%) but AMiM stays ≤ 0.20; raw attention improves with fine-tuning (PG 7.2% $\rightarrow$ 42.9%) but AMiM saturates at 0.214; rollout and Winsor-CAM sit at or below the random-mask baseline, and the Qwen3-VL probe is weakest (PG exactly 0%). *In contrast*, both CXR-trained classifiers produce $p<10^{-4}$ gaps over shuffled on every metric, so the negative VLM result is not an artifact of the imaging modality or the localization metric.

4.2 Counterfactual Perturbations

Table 3 summarises patch-occlusion causal grounding across the MedGemma variants and cross-family probes on PadChest. Spearman ρ between raw-attention saliency and the causal-importance map is negative for every MedGemma variant (-0.098 base, -0.050 targeted LoRA, -0.168 full LoRA at n=500), persisting at the n=12K multi-task scale-up. Qwen3-VL agrees in direction ($\rho=-0.117$ on its native 8×8 grid, n=637). LLaVA-RAD reverses sign ($\rho=+0.137$), but its causal AMiM is close to zero (0.005 vs. 0.066 attention): it barely uses the image,

Table 2. Attribution method comparison on CheXlocalize (n=643). PG: pointing-game hit rate (%). AMiM: attribution mass in mask. Pixel-AP: pixel average precision. Δ: gap between true and shuffled-mask saliency (positive favors true). Best VLM Δ per metric is **bold**. Among VLMs no method exceeds 22% AMiM; rollout and Winsor-CAM give $\Delta \approx 0$. Both CXR classifiers produce larger gaps on the same images and classes ($p<10^{-4}$).

		PG (%)		AMiM		Pixel-AP	
Method	Model	true	Δ	true	Δ	true	Δ
Raw attention	Base	7.2	+3.9	0.158	+0.043	0.211	+0.073
	Targeted LoRA	21.0	+12.0	0.189	+0.059	0.264	+0.107
	Full LoRA	42.9	+22.6	0.214	**+0.066**	0.340	**+0.149**
Attention rollout	all 3 MedGemma	0.3	+0.2	0.073	≈ 0	0.068	≈ 0
Gradient attention	Base	44.3	+20.8	0.178	+0.044	0.296	+0.113
	Targeted LoRA	39.7	+17.1	0.176	+0.044	0.279	+0.103
	Full LoRA	50.9	**+26.4**	0.200	+0.057	0.333	+0.139
Winsor-CAM	Base	12.0	+2.2	0.109	+0.004	0.143	+0.018
	Targeted LoRA	7.9	−0.2	0.099	+0.002	0.120	+0.010
	Full LoRA	9.8	+3.1	0.103	+0.006	0.126	+0.015
Cross-architecture probes:							
Raw attention	LLaVA-RAD	0.2	−0.2	0.076	+0.015	0.103	+0.017
Raw attention	Qwen3-VL-8B	0.0	+0.0	0.063	+0.005	0.081	+0.006
Localization-only baselines (CXR CNN classifiers, n=536 matched findings):							
Grad-CAM	DenseNet-CXR-Chex	37.5	+12.5	0.209	+0.046	0.306	+0.074
Grad-CAM	ResNet50-CXR-All	28.7	+13.6	0.170	+0.043	0.213	+0.060

so its attention agrees with the very small causal signal that exists, not clinical anatomy. For the three MedGemma variants the AMiM difference (causal − attention) is positive and significant ($p_{\text{perm}}<0.001$, n=474), largest for full LoRA, the variant with the lowest flip rate: consistency-improving fine-tuning does not align attention with cause. Restricting to the n=474 images all five models processed preserves the pattern, with all 95% CIs excluding zero.

The mismatch is not specific to raw attention (Table 3, lower panel): rollout, gradient×attention, and the axiomatic integrated gradients [27] all anti-correlate with causal importance for every MedGemma variant (−0.18 to −0.28 against the absolute-shift causal map, n=474, CIs exclude zero), Winsor-CAM is near zero, and the model-agnostic RISE [25] is near-uniform (ρ between +0.009 and +0.010, stable to N=1500 masks) because the models barely respond to random masking.

Second annotated dataset (VinDr-CXR). The grounding failure replicates on VinDr-CXR [28] (n=300 presence questions with radiologist boxes per model). Attention mass in the radiologist box stays low (0.075–0.105), below the causal map (0.100–0.133), and true-box AMiM exceeds shuffled controls on only 27–33% of cases, so attention again fails to concentrate on annotated anatomy. The attention-causal correlation is near zero (ρ=+0.061 base, +0.061 targeted LoRA, +0.053 full LoRA), so attention does not track causal importance on a second dataset either; the residual correlation is dataset-dependent in sign

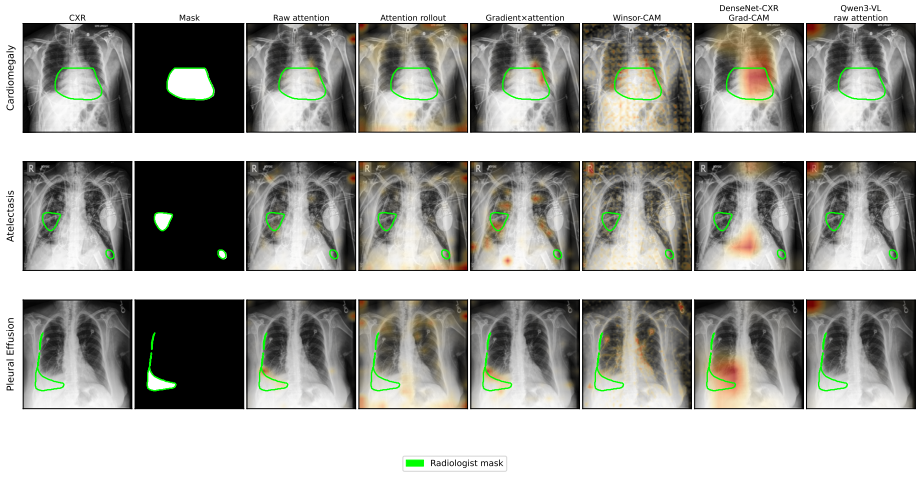

Fig. 2. Qualitative attribution comparison on three CheXlocalize cases. Columns show the CXR, radiologist mask, MedGemma-4B base methods, DenseNet-CXR Grad-CAM, and Qwen3-VL raw attention. MedGemma maps are broad or near-empty and Qwen3-VL is nearly uniform; DenseNet-CXR is the only method whose peak reliably falls near the mask.

($|\rho| < 0.17$ throughout), itself a reason not to read heatmap-causal alignment as reliable.

CXR-specialist VLM (CheXagent): grounding is dataset-dependent. CheXagent-2-3b [29] does not expose cross-modal attention, so we probe its decision with the margin-only ROI test (occlude the radiologist box vs. a random same-size region). It is not grounded on PadChest (+0.038, 95% CI [−0.025, +0.100]) but *is* on VinDr-CXR (+0.568 gray, +0.692 blur); the effect strengthens under blur, ruling out a gray-opacity artifact. A specialist VLM can be causally grounded on the annotation style it matches but not in general.

Insertion/deletion fidelity. On the standard insertion/deletion protocol (n=200 per VLM), DenseNet-CXR shows a real saliency effect (insertion AUC 0.690 vs. random 0.666, Δ=+0.024, $p_{\text{perm}}<10^{-4}$) while all five VLMs have uniformly small effects ($|\Delta\text{AUC}| \leq 0.024$). The VLM methods fail all three probes (annotation alignment, input-attribution alignment, the cross-architecture sanity check) in the same direction while DenseNet passes the annotation and insertion probes, so the result holds across independent axes rather than one metric.

Table 3. Patch-occlusion causal grounding on PadChest. *Top:* raw-attention Spearman ρ vs. the signed causal map; AMiM Δ is causal minus attention mass inside the radiologist bbox ($p_{\text{perm}}<0.001$ where shown; 95% CIs for ρ). *Bottom:* per-method ρ vs. absolute margin shift, explaining why raw attention differs from the signed-ρ value. Negative ρ persists under the 24× scale-up.

Model	ρ	95% CI	L_1	Causal	Attn.	Δ
MedGemma family (n=474; 16×16 grid):						
Base	−0.098	[−0.121,−0.075]	1.440	0.148	0.087	+0.061
Targeted LoRA (500)	−0.050	[−0.071,−0.030]	1.442	0.147	0.085	+0.062
Targeted LoRA (12K MT)	−0.115	[−0.136,−0.095]	1.388	0.160	0.092	+0.068
Full LoRA (500)	−0.168	[−0.190,−0.146]	1.433	0.180	0.091	+0.088
Full LoRA (12K MT)	−0.145	[−0.167,−0.123]	1.443	0.162	0.085	+0.077
Cross-family probes (n=637; native grid):						
Qwen3-VL-8B (8×8)	−0.117	[−0.130,−0.104]	1.481	0.031	0.198	−0.167
LLaVA-RAD (16×16)	+0.137	[+0.117,+0.156]	1.276	0.005	0.066	−0.062

Per-method ρ vs. the absolute-shift causal map (n=474; RISE n=200, N=320):

Method	Base	Targeted LoRA	Full LoRA
Raw attention	−0.263	−0.202	−0.253
Attention rollout	−0.257	−0.178	−0.238
Gradient × attention	−0.243	−0.189	−0.258
Winsor-CAM	−0.059	+0.052	+0.040
Integrated gradients	−0.276	−0.208	−0.281
RISE	+0.010	+0.009	+0.009

5 Discussion and Conclusion

Consistency without grounding. Reducing paraphrase flip rates does not produce models that look at the right anatomy. Fine-tuning trades consistency for text-only agreement and yields only modest grounding gains: raw-attention AMiM still leaves four fifths of mass outside the mask, and Winsor-CAM does not improve. Consistency tells you about the mechanism, not whether the model reads the evidence.

No evaluated VLM has grounded attention. Grounding needs two conditions together (Fig. 3): the MedGemma variants and Qwen3-VL use the image but their attention fails to track cause (negative Spearman ρ, surviving the n=12K multi-task scale-up), while LLaVA-RAD correlates positively only because it fails the first condition (causal attribution mass 0.005), so its ρ joins two near-zero signals. CheXagent shows a decision can still be causally grounded in-distribution (VinDr), so we document that heatmaps do not certify grounding, not that medical VLMs never use the right region. The failure is not method- or modality-specific: no VLM method exceeds 22% mask AMiM, and two CXR-trained classifiers pass the same protocol. Whether architectures with explicit grounding heads such as MIMO [16] fare better remains open.

Scope and limitations. The DenseNet/ResNet contrast does not claim classifiers are globally better; it shows the protocol detects localized saliency when a model is trained for CXR classification. Causal maps use the 16×16 token grid, so sub-patch pathologies may be partly masked, and Qwen3-VL is

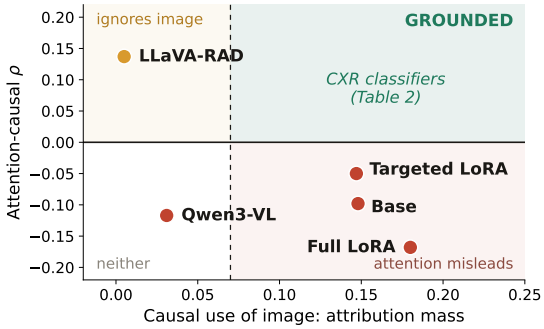

Fig. 3. Grounding needs both image use (horizontal: causal attribution mass) and attention-causal alignment (vertical: Spearman ρ). Points are exact Table 3 values; the dashed line is a visual guide. No VLM reaches the grounded quadrant; the CXR classifier controls are grounded by Table 2.

evaluated at 8×8 for memory. A fill-sensitivity check yields no positive ρ under any fill, so the reported value is a lower bound.

Conclusion. Heatmaps from these medical VLMs are reassuring but not faithful: across bounding-box overlap, pixel-mask AMiM, and patch-occlusion Spearman ρ, attention fails where the CXR classifier baselines pass. Clinical explanations should be validated by controlled localization metrics and causal perturbation, not by visual inspection alone.

Use of AI tools. We used a large language model (Claude Opus 4.8) for language editing and LaTeX formatting (grammar, sentence-level rewriting, layout); it did not generate research questions, methods, experiments, or analysis. The authors verified all numbers, citations, and claims and take full responsibility.

Acknowledgments. Conducted at the Secure and Assured Intelligent Learning (SAIL) Lab, University of New Haven. No specific external funding was received.

Disclosure of Interests. B. Sadanandan is a research engineer at Medtronic Medical Surgical, CT, USA; this employment is unrelated to the present work. The authors cite related primary studies of their own [17,18,19], published or under review elsewhere. No other competing interests are declared.

References

1. Zambrano Chaves, J.M., et al.: Towards a clinically accessible radiology foundation model: open-access and lightweight, with automated evaluation. arXiv:2403.08002 (2024)
2. Hu, E.J., et al.: LoRA: low-rank adaptation of large language models. In: ICLR (2022)
3. Ribeiro, M.T., Wu, T., Guestrin, C., Singh, S.: Beyond accuracy: behavioral testing of NLP models with CheckList. In: ACL, pp. 4902–4912 (2020)
4. Bustos, A., Pertusa, A., Salinas, J.-M., de la Iglesia-Vayá, M.: PadChest: a large chest X-ray image dataset with multi-label annotated reports. Med. Image Anal. **66**, 101797 (2020)
5. Johnson, A.E.W., et al.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci. Data **6**(1), 317 (2019)

6. Lieberum, T., et al.: Gemma Scope: open sparse autoencoders everywhere all at once on Gemma 2. arXiv:2408.05147 (2024)
7. Chefer, H., Gur, S., Wolf, L.: Transformer interpretability beyond attention visualization. In: CVPR, pp. 782–791 (2021)
8. Jain, S., Wallace, B.C.: Attention is not explanation. In: NAACL, pp. 3543–3556 (2019)
9. Wiegreffe, S., Pinter, Y.: Attention is not not explanation. In: EMNLP, pp. 11–20 (2019)
10. Xia, P., et al.: CARES: a comprehensive benchmark of trustworthiness in medical vision language models. arXiv:2406.06007 (2024)
11. Gu, Z., Yin, C., Liu, F., Zhang, P.: MedVH: towards systematic evaluation of hallucination for large vision language models in the medical context. arXiv:2407.02730 (2024)
12. Saporta, A., et al.: Benchmarking saliency methods for chest X-ray interpretation. Nat. Mach. Intell. **4**(10), 867–878 (2022)
13. Abnar, S., Zuidema, W.: Quantifying attention flow in transformers. In: ACL, pp. 4190–4197 (2020)
14. Wan, X., et al.: Eliminating language bias for medical visual question answering with counterfactual contrastive training. In: MICCAI (2025)
15. Han, T., Adams, L.C., Nebelung, S., et al.: Multimodal large language models are generalist medical image interpreters. medRxiv 2023.12.21.23300146 (2023)
16. Chen, Y., Xu, D., Huang, Y., et al.: MIMO: a medical vision language model with visual referring multimodal input and pixel grounding multimodal output. arXiv:2510.10011 (2025)
17. Sadanandan, B., Behzadan, V.: PSF-Med: measuring and explaining paraphrase sensitivity in medical vision language models. arXiv:2602.21428 (2026)
18. Sadanandan, B., Behzadan, V.: Mechanistically guided LoRA improves paraphrase consistency in medical vision-language models. In: Proceedings of Machine Learning Research (CHIL 2026), vol. 333, pp. 703–720 (2026)
19. Sadanandan, B., Behzadan, V.: Predictive entropy links calibration and paraphrase sensitivity in medical vision-language models. arXiv:2604.08941 (2026)
20. Alsentzer, E., et al.: Publicly available clinical BERT embeddings. In: Clinical NLP Workshop, NAACL, pp. 72–78 (2019)
21. Sellergren, A., Kazemzadeh, S., Jaroensri, T., et al.: MedGemma technical report. arXiv:2507.05201 (2025)
22. Selvaraju, R.R., et al.: Grad-CAM: visual explanations from deep networks via gradient-based localization. In: ICCV, pp. 618–626 (2017)
23. Cohen, J.P., Viviano, J.D., Bertin, P., et al.: TorchXRayVision: a library of chest X-ray datasets and models. In: MIDL (2022)
24. Bai, S., Cai, Y., et al.: Qwen3-VL technical report. arXiv:2511.21631 (2025)
25. Petsiuk, V., Das, A., Saenko, K.: RISE: randomized input sampling for explanation of black-box models. In: BMVC (2018)
26. Adebayo, J., et al.: Sanity checks for saliency maps. In: NeurIPS, vol. 31 (2018)
27. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: ICML, pp. 3319–3328 (2017)
28. Nguyen, H.Q., et al.: VinDr-CXR: an open dataset of chest X-rays with radiologist's annotations. Sci. Data **9**(1), 429 (2022)
29. Chen, Z., Varma, M., Delbrouck, J.-B., et al.: CheXagent: towards a foundation model for chest X-ray interpretation. arXiv:2401.12208 (2024)
30. Lan, Z., Sun, L., Walter, M.R., Zhou, J.: Seeing without looking: do vision-language benchmarks really test vision? arXiv:2605.22903 (2026)
31. Aranya, O.F.M.R.R., Desai, K.: To agree or to be right? the grounding-sycophancy tradeoff in medical vision-language models. arXiv:2603.22623 (2026)
32. Mayer, L., et al.: 6 fingers, 1 kidney: natural adversarial medical images reveal critical weaknesses of vision-language models. arXiv:2512.04238 (2025)