

Supplementary Material

How Well Do Chest X-Ray VLM Attention Overlays Match Radiologist Boxes? A Cross-Model Audit and Radiologist Reader Study

Syed Faizan¹, Mahesh Vasamsetti¹, Kotak Shivam Dilipbhai², Gundamaraju
Madhavi Latha³, and Parameswari Krishnamurthy¹

¹ Language Technologies Research Centre, IIIT Hyderabad, India

² Parul University, Vadodara, Gujarat, India

³ Independent Radiologist, India

This supplement provides the attention-extraction pipeline, the full robustness-control battery, extended qualitative evidence, and protocol details for the main paper. It is not required to follow the main results; section, table and figure numbers prefixed “S” are supplementary, and references to sections such as “§4.3” point to the main paper.

1 Attention-extraction pipeline

Figure 1 illustrates the per-case pipeline summarised in §3 of the main paper. The image and a class-conditional prompt (e.g. “*There is Pneumonia visible in this chest radiograph.*”) pass through the frozen VLM and the cross-attention tensor of shape $(B=1, H, q, k)$ is recorded at the chosen layer, where H is the number of heads and q, k index the query (text) and key (image-or-text) token positions. We keep the assistant-content token rows (dropping punctuation, stopwords, and special tokens) and the image-patch columns, average over heads and retained text tokens, and reshape onto the model’s native patch grid; the grid is L_1 -normalised so its values sum to one and read as a spatial probability map.

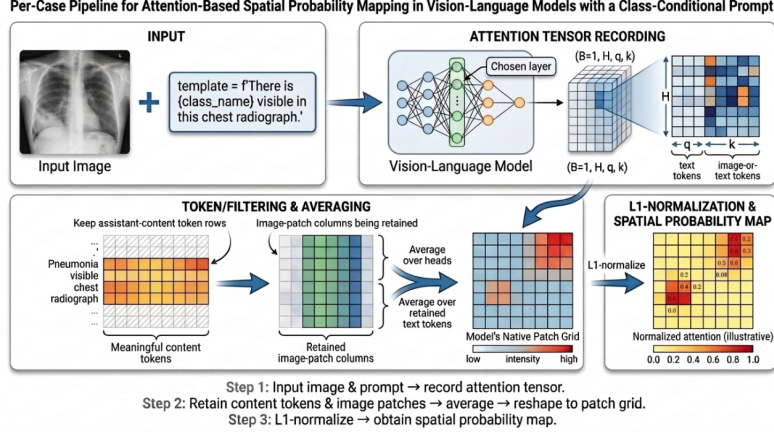


Fig. 1. Per-case attention-extraction pipeline (§1).

2 Robustness controls

Because absolute alignment AUC mixes per-image localisation with several confounds, we report a battery of controls (Table 1) that isolate the four most plausible sources of inflation: anatomical prior, deployment regime, attribution channel, and head aggregator.

Table 1. Per-checkpoint summary on RSNA Pneumonia ($n=500$). “Own-prior gap” is per-case AUC minus the model’s image-invariant LOO-mean; “black-img cos” is the real-vs-black-image cosine; “free-gen AUC” is deployment-time (non-emission at 0.5), emission-conditional in parentheses.

Model	AUC	own-prior gap	black cos	free-gen AUC
MAIRA-2	0.896	+0.076 *	0.13	0.793 (0.871)
MedGemma	0.748	−0.042 *	0.49	0.597 (0.725)
LLaVA-Med	0.605	+0.016	0.13	0.500 (n/a)
LLaVA-1.5	0.594	0.000 [‡]	0.04	0.500 (n/a)
Grad-CAM DenseNet-121 [†]	0.539	n/a	n/a	n/a
Dataset-mean-bbox oracle	0.887	n/a	n/a	n/a

*Significant by paired bootstrap ($n=500$, 10,000 resamples; survives Bonferroni $k=20$). [‡]LLaVA-1.5’s gap is +0.0003, shown as 0.000. [†]DenseNet-121 (torchxrayvision); Grad-CAM on the Pneumonia logit, same per-cell AUC (486/500 valid).

- *Own-prior gap (two baselines).* Against each model’s image-invariant LOO-mean (a test of per-image *modulation*) only MAIRA-2 gains (+0.076); MedGemma

falls below its own mean (-0.042) and the LLaVA pair match theirs. But against the class-conditional atlas (oracle 0.887) even MAIRA-2 only ties the prior (RSNA $+0.008$) and falls below it on VinDr (-0.058 ; main paper §4.3): no model beats anatomy on typical cases.

- *Non-VLM bounds (RSNA)*. A DenseNet-121 Grad-CAM (0.539) and a bbox-supervised BasicUNet (0.940) bracket MAIRA-2’s zero-shot 0.896, which still does not exceed the anatomical prior (main paper §4.3).
- *Deployment regime*. In free generation a model yields a scorable overlay only when it names the finding (the *emission rate*); the *intention-to-treat* (ITT) AUC scores non-emitting cases at chance, unlike the emission-conditional AUC over cases that named it. Emission rates are MAIRA-2 75/18/7%, MedGemma 38/10/9%, LLaVA pair 0% (RSNA/MS-CXR/VinDr); emission-conditional pooled AUC is 0.871 (MAIRA-2) and 0.725 (MedGemma), ITT 0.610 and 0.548. MedGemma’s overlay is timing-locked, not position-invariant like MAIRA-2’s: its free-generation AUC is 0.416 at the first token but 0.798 at the disease-mention token ($\Delta = +0.373$, $p < 10^{-23}$), so an overlay must be read at the finding token, not the first token most prototypes display.
- *Attribution and head aggregator*. The ranking survives raw attention, attention $\times$ gradient, input-gradient, Integrated Gradients, and attention rollout (Abnar and Zuidema, ACL 2020); a fixed-head AUC lifts MedGemma to 0.814 and LLaVA-1.5 to 0.717 without changing the order (LLaVA-1.5’s gradient saliencies hit chance despite $\|\nabla\|_\infty \approx 12$ in fp32).

3 Image content versus prompt and position

A high alignment AUC need not mean a model is *looking at the image*: it could attend the lung fields by default, as a learned anatomical bias. Replacing the image with a content-free version (prompt fixed), an all-black image reconstructs cosine 0.49 of MedGemma’s map but only 0.13 of MAIRA-2’s (Table 1, Fig. 2); a horizontal flip (cosine 0.972 vs 0.611) and six further perturbation controls rule out a vision-encoder artefact, and three paraphrases change AUC by $|\Delta\text{AUC}| < 0.013$. Scoring content-free attention against the boxes is at chance (MAIRA-2 0.52, MedGemma 0.34), so the 0.77–0.88 alignment is image-driven, not centre bias.

Mean cross-attention on real CXR vs an entirely black image
(cosine of the averaged maps; same prompt), all four checkpoints.

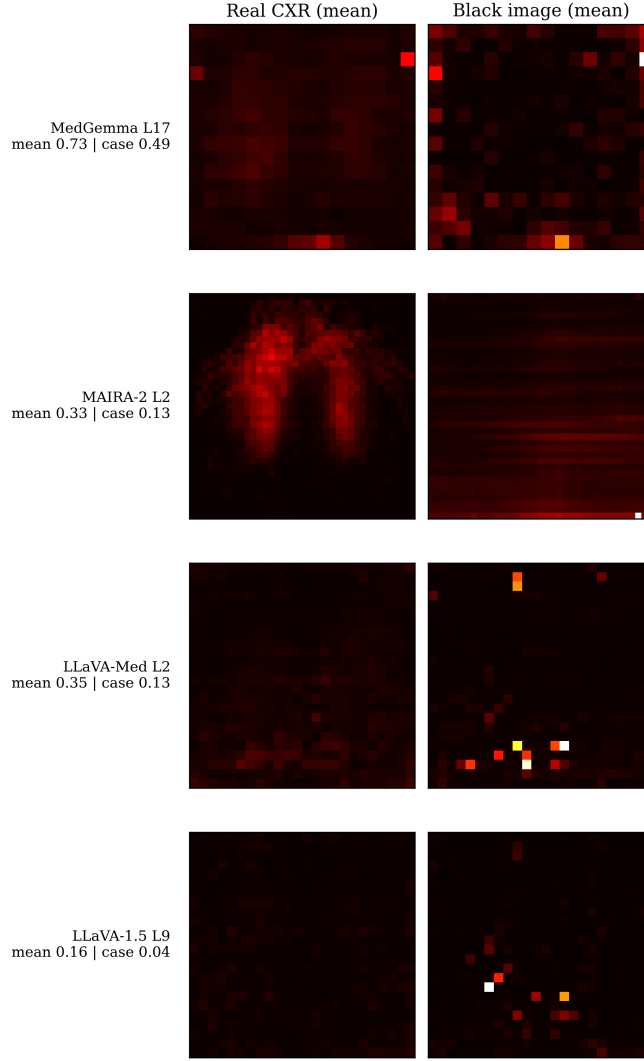


Fig. 2. Mean cross-attention on the real CXR (left) versus an entirely black image (right) under the same prompt, for all four checkpoints at their selected layers. Per-case cosines match the “black-ing cos” column of Table 1.

4 Additional qualitative overlays

Figure 3 extends Fig. 1 of the main paper with a second representative case. Both cases are drawn from the radiologist reader study and are shown de-anonymised

(the two readers scored them shuffled and unlabelled). For each case the left panel draws the radiologist bounding box (green) and the remaining panels are the exact attention heat-maps the readers scored, annotated with each model’s per-case Alignment-AUC. In both the Pneumothorax and the Calcification case MAIRA-2 concentrates on or near the annotated finding and attains the highest per-case AUC, whereas MedGemma and the two LLaVA checkpoints remain diffuse or off-target, and the class-conditional anatomical prior is a broad lung-field pattern that still partially overlaps the box.

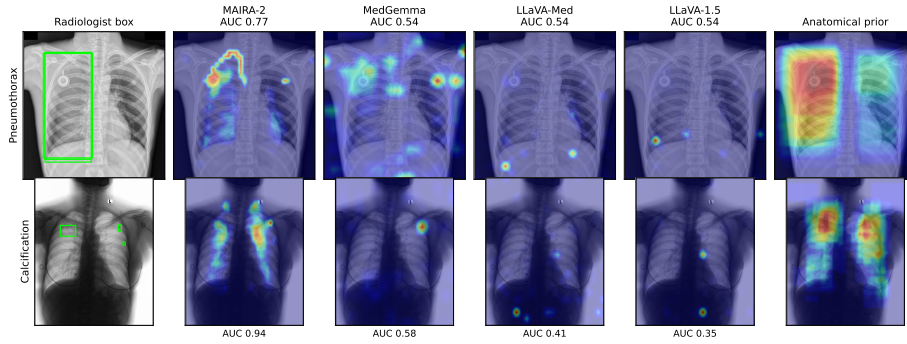


Fig. 3. Two representative reader-study cases (top: VinDr Pneumothorax; bottom: VinDr Calcification), de-anonymised. Left: radiologist box (green). Remaining columns: per-model attention overlays with per-case Alignment-AUC, and the anatomical prior (right).

5 Layer-selection protocol

For each model the single analysed layer ℓ^* is the argmax, over decoder layers, of the mean per-case Alignment-AUC on a held-out 200-case multi-finding VinDr development split disjoint from every test set (§3 of the main paper). This yields L2 for MAIRA-2, L17 for MedGemma, L2 for LLaVA-Med, and L9 for LLaVA-1.5 (main-paper Table 1). To show that the reported cross-model ranking does not depend on a single lucky choice, we run a 250/250 held-out protocol on the independent RSNA cohort: over 200 random half-splits we re-select ℓ^* on one half and re-evaluate it on the other, which propagates the layer-selection variance into the ranking. The ordering MAIRA-2 $\succ$ MedGemma $\succ$ LLaVA pair is preserved across these resamples. As noted in the main-paper limitations, this protocol establishes stability of the cross-model *ranking*, not full robustness to layer choice within a single model.

6 Reader-study protocol

Two board-certified radiologists independently scored all 124 reader-study cases. Each case presented six overlays: the four model overlays, the class-conditional anatomical prior, and a random-noise control, rendered identically and shuffled to a different order per case, so the readers saw only anonymised panels A–F. Each overlay was rated 1–5 for how well the highlight matched the named finding, on the rubric 1 = wrong / unrelated region, 2 = mostly off-target, 3 = partially overlaps the finding, 4 = correct region but slightly broad, 5 = tight and correct. For every case the readers also flagged the best, worst, and any misleading overlays, and judged whether the dataset bounding box was clinically acceptable and whether it captured the finding. Aggregated results (mean plausibility, best/worst/misleading shares, and the Alignment-AUC correlation) are reported in Table 4 and §4.5 of the main paper.