STEERABLE BUT NOT READABLE: FIXATION HEADS IN MEDICAL VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical vision-language models (VLMs) write fluent findings but do not show which image region supports each sentence, so a clinician cannot check a claim against the scan. The usual remedy reads an attention or attribution map over the scan as a record of what the model used. Generalist VLMs contain a few attention heads that follow the region being described, yet it is unknown whether medical fine-tuning keeps these heads and whether their attention can serve as provenance. We test these *fixation heads* in medical VLMs in both directions, reading their attention as evidence and writing a region into it. Our evaluation across three model families and six corpora shows that (i) every model has fixation heads and supervised medical tunes inherit them from their base, (ii) the heads are steerable, as aiming them at a region changes what the model says, and (iii) they are not readable, as their attention barely follows the question, leaves the organ as the model answers and locates findings no better than random boxes, as do standard attribution maps. Anatomy-conditioned generation turns the write into provenance without training, writing one anatomical region into the heads per sentence so that the sentence depends on it, and adds clinical findings to chest reports. MedGaze, a low-rank adapter on the heads' query rows, learns the write, moving the heads onto the anatomy with no region at test and answering organ questions more accurately than training-free decoders. Our findings emphasize the need to write provenance during decoding rather than read it from attention maps.

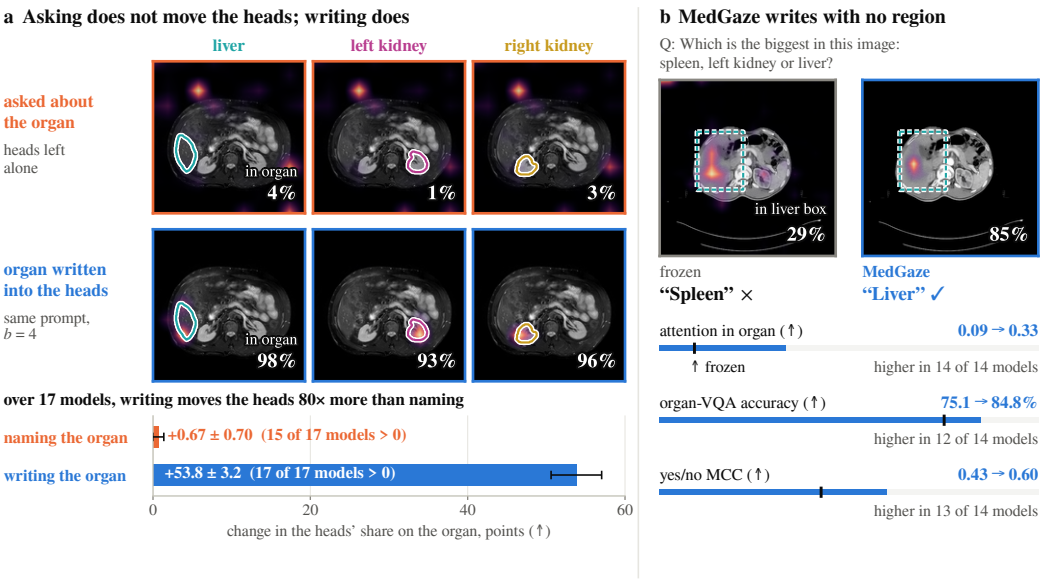


Figure 1: **Fixation heads can be written but not read.** (a) Asked about three organs of an MRI, the fixation heads of `Lingshu-7B` put 1 to 4% of their attention in the organ, and 93 to 98% when it is written in. (b) With no region, MedGaze fixes the answer.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Clinicians need medical vision-language models (VLMs) to show *provenance*, the image region that supports each claim the model makes. A medical VLM asked about a scan gives a fluent, clinical answer, but it does not say where in the image the answer came from. In Figure 1b, a model asked which organ is biggest answers "spleen" although the liver is, and nothing in the answer tells a reader which region it relied on. Provenance would (i) let a radiologist check each claim against the scan and (ii) let an auditor trace a report sentence to the region it describes. The need is pressing because audits find that chest radiograph models often ignore the image (Lotfinia et al., 2026), report generators exploit prevalence shortcuts (Ruffini et al., 2026), and surgical question answering shows the same pattern (Shin et al., 2026).

Provenance is hard to obtain because a VLM never states which image tokens its answer used, so it has to be inferred from the model's internals. Attention and attribution maps infer it after the fact, which assumes that the map records what the model used (Abnar & Zuidema, 2020; Chefer et al., 2021). Behavioural audits show that the image is under-used but do not locate the component responsible. In generalist VLMs, Gandikota & Bau (2026) found such a component, a few attention heads that track the region being described, so that redirecting them redirects the description. Medical VLMs are fine-tuned from these backbones, and heads that route regions are an opportunity, as they could either report the region a claim came from or impose it.

This paper studies these heads in medical VLMs under the name *fixation heads*, because they fixate on a named region, whereas "gaze" would suggest that they reveal where the model looks. To *read* the heads is to treat their attention as evidence of what the model used, and to *write* a region is to bias their attention toward it, with *steering*, a fixed bias on their attention logits, as the simplest write. The key observation is that the heads can be written but not read. Asking `Lingshu-7B` about the liver or either kidney of one scan barely moves its fixation heads, whereas writing the same organ into the same heads puts nearly all of their attention inside it, as Figure 1a shows, and over 17 models writing at $b = 4$ moves the heads 80 times more than asking. Provenance should therefore be written during decoding, and *anatomy-conditioned generation* (ACG) writes one anatomical region into the heads per sentence, so every sentence carries the region it was decoded under. *MedGaze*, named for setting where the heads look, learns the same write with a low-rank update of only the heads' query rows, trained with an attention loss on the organ boxes of the training images.

The evaluation covers 25 VLMs from three families, 2B to 73B parameters, including 14 medical tunes paired with their own bases, on six corpora of radiographs, CT, MRI and histopathology. Fixation heads exist in every model, and steering them redirects what the model describes more often than random heads in 14 of 17 models. Reading them fails, as their attention barely follows the question, and their share on the organ rises to 8.5% while 7 models reason but falls to 4.3% on the answer word, against 6.0% chance. Writing succeeds, as a steered *region tour* of the chest raises clinical finding F1 by $+7.85$ points, and blanking the written region costs the steered tour 9.2 more F1 points than the unsteered one. MedGaze raises organ question accuracy from 75.1 to 84.8%.

In this paper, our major contributions are:

1. We show that **fixation heads are steerable but not readable**, so their attention is a control channel, not an evidence channel. Writing a region into the heads moves their attention onto it 80 times more than asking about it ($b = 4$), yet their attention fails all five tests of reading, falling below chance on the answer word (4.3% against 6.0%), and standard attribution maps do no better than random boxes (§5.2, §5.3).
2. We propose **Anatomy-conditioned generation (ACG)**, a training-free method that ties every sentence to one anatomical region written into the heads. Where head steering redirects a whole description, ACG writes a label-free region per sentence, and blanking it costs the steered report tour 9.2 more F1 points than the unsteered one (§5.4).
3. We introduce **MedGaze**, a low-rank adapter that trains the fixation heads to find the relevant anatomy themselves, so no region is needed at test time. Unlike adapters that tune whole layers, it updates only the fixation heads' query rows and only while the answer is decoded, 0.45M to 1.00M parameters that leave every other head and the prompt's cache unchanged. It raises attention on the organ from 0.09 to 0.33 and accuracy from 75.1% to 84.8%, and is more accurate than VCD, PAI and CRG (§5.5).

4. **A head-level comparison of 14 medical tunes with their own bases**, showing that supervised fine-tuning inherits region routing: all 4 supervised Qwen tunes keep their base's top head, and inherited heads keep their fixation matrix ($r = 0.91$; §5.1).

2 Related Work

Attention heads in VLMs. A key challenge in interpretability is to tie a behaviour to a testable component, and head-level studies meet it by isolating heads with one job, such as localisation heads (Kang et al., 2025a), retrieval heads (Wu et al., 2025), truthfulness heads (Li et al., 2023) and visual attention sinks (Kang et al., 2025b), and toolkits carry probing and patching over to VLMs (Joseph et al., 2025; Zheng et al., 2024). Gandikota & Bau (2026) find and steer region-following heads on natural images, and our score follows theirs; PASTA steers selected heads of language models by rescaling their attention (Zhang et al., 2024b). Whether attention explains a prediction has been contested in text models (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019; Serrano & Smith, 2019). Our work tests inheritance against each tune's own base, reads and writes the same heads, and turns writing into provenance.

Medical VLMs and whether they use the image. Clinical use depends on whether a medical VLM grounds its answer in the image. Open medical VLMs adapt public generalist backbones, as Lingshu (Xu et al., 2026) and HuatuoGPT-Vision (Chen et al., 2024) build on Qwen2.5-VL (Bai et al., 2025a), MedGemma builds on Gemma3 (Sellergren et al., 2025), and reasoning tunes add reinforcement learning (Huang et al., 2025; Mullappilly et al., 2026; Pan et al., 2025). Audits find that such models often ignore the image (Lotfinia et al., 2026; Ruffini et al., 2026; Shin et al., 2026), that benchmarks can be solved without looking (Lan et al., 2026; Singla et al., 2026), and that models can attend to the right region yet answer wrongly (Liu et al., 2026). Grounded corpora such as MS-CXR (Boecking et al., 2022; Johnson et al., 2019), VinDr-CXR, Chest ImaGenome and SLAKE supply radiologist boxes and organ masks, and a chest anatomy segmenter (Cohen et al., 2022) and Segment Anything (Kirillov et al., 2023; Ma et al., 2024) supply regions without labels, but all of them score grounding after the fact. Our work uses the same corpora and segmenters to supply the regions that are written into the heads during decoding.

3 Method: Finding and Writing Fixation Heads

The method (a) finds the heads that follow a named region, (b) writes a region into them by steering and ACG, and (c) learns the write with MedGaze, as the panels of Figure 2 show.

3.1 Discovering fixation heads

Finding the heads requires a score that separates a head that follows the named region from one that attends to a fixed place. Take an image with N named regions $R_1, \ldots, R_N$ and one query q_i per region that names R_i. One prefill per query gives the attention row of the *final prompt token*, the position whose output predicts the first answer token. For head h, $A_h(i, j)$ is the attention that this row places on the image tokens of R_j under query q_i. This $N \times N$ *fixation matrix* is diagonal for a head that looks at whatever region the query names, and the fixation score is its mean diagonal, $s_h = \frac{1}{N} \sum_i A_h(i, i)$, as Figure 2a illustrates for the 784 heads of `Lingshu-7B`. Attention is summed raw and never renormalised across regions, so a head that ignores the image scores zero instead of looking perfectly diagonal. Renormalising each row over the N regions gives a head's *concentration*, with chance level $1/N$, which compares models on one scale. Every other result reports a head set's *share* on a region, its attention on the region's image tokens divided by its attention on all image tokens, averaged over the heads. The top 100 heads form the fixation set $\mathcal{H}$, and 100 heads drawn with a fixed seed from outside $\mathcal{H}$, the *random heads*, are the baseline in every experiment.

3.2 Region protocols

The score needs regions that a query can name without revealing their content. **Montage.** Six unrelated medical images are stitched into a strip and queried by position ("the k-th image"), so the

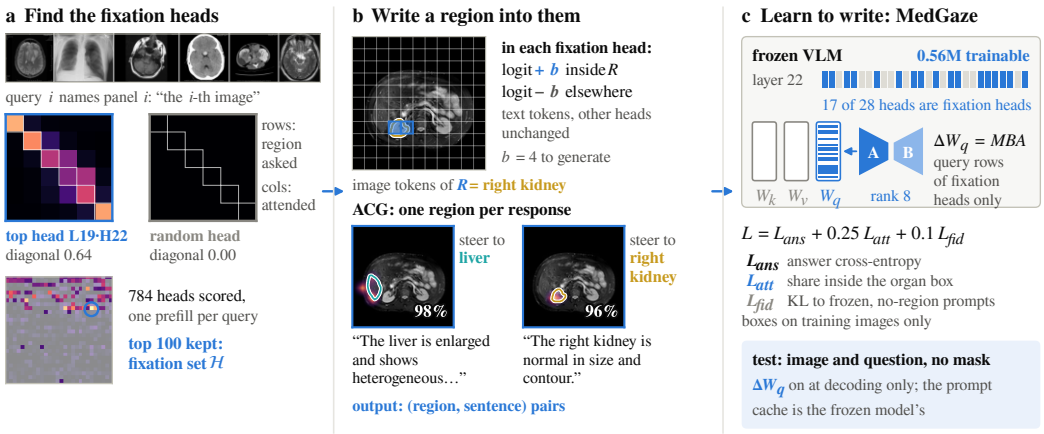


Figure 2: **Discovery finds the heads that follow a named region, and both methods write a region into them.** (a) Discovery on `Lingshu-7B` (grid: layers × heads, coloured by score); (b) steering and ACG; (c) MedGaze, which needs no region at test.

ground truth holds by construction. **Radiologist boxes.** Heads are also scored inside radiologist boxes against equal-area control boxes, which Appendix B.2 defines, and heads ranked this way are *box-ranked*. **Segmented anatomy.** Regions come from anatomical masks on one unmodified image, the lungs, heart and spine of a chest anatomy segmenter (Cohen et al., 2022) on radiographs and the corpus's organ masks on CT and MRI. The query names the structure, and a token belongs to a region when at least a quarter of its patch lies inside the mask, which Appendix B.2 details.

3.3 Steering

Testing whether the heads cause what the model says requires an intervention on their attention alone. Steering biases their attention logits before the softmax, adding $+b$ for each image token $t \in R$ and $-b$ for every other image token, for each $h \in \mathcal{H}$, as Figure 2b shows. Text tokens and all heads outside $\mathcal{H}$ keep their original logits, so the intervention changes only where the fixation heads look. Generation uses $b = 4$, the largest value that keeps text clean in the bias sweep of Appendix Table 4, whereas the montage test follows the reference recipe's hard gate $b = 10^4$, as does scoring a fixed phrase. Because the first answer token is predicted from the last prompt position, the decoder prefills all but that token unsteered and runs it as a steered step.

3.4 Anatomy-conditioned generation

Steering needs a region, and ACG turns it into a procedure that needs no training and changes neither the prompt nor the image. A region source proposes regions $R_1, \ldots, R_K$, and for each region ACG steers $\mathcal{H}$ to it at $b = 4$ and generates one response. The output is a list of (region, sentence) pairs, so each claim carries the region written into the heads while it was decoded, which records provenance rather than certifying that the sentence is correct. Four region sources run on identical images, models and heads, a 3×3 grid, the annotated box as an oracle, organ or anatomy masks, and Segment Anything in automatic mode. To score localisation, the selected region is the one under whose steering the reference phrase is most likely, and a *hit* means that this region contains the annotated box.

3.5 Learning to write: the MedGaze adapter

ACG needs a region at test time, as does calibrated fixation steering, which fits one gain per head in place of b (Algorithm 1, Appendix B.5), and MedGaze removes this need by training the heads to look at the anatomy themselves, with the backbone frozen. MedGaze adds a low-rank update to the query projection of the fixation heads only, as Figure 2c shows. Updating only the queries changes *where* the heads look, whereas the keys and values, and hence *what* the heads read, stay those of the

base model,

$$q = W_q x + \tfrac{\alpha}{r}\, MBAx. \tag{1}$$

Here W_q is the frozen query weight, A and B are rank-r factors with B initialised to zero, and M writes BA into the fixation heads' rows only, so every other head stays bit-identical to the base model. The update is off during prefill and on at the answer step, so the key-value cache is the frozen model's and the adapter can be switched off without recomputing the prompt. The training loss is

$$\mathcal{L} = \mathcal{L}_{\text{ans}} + \lambda_{\text{att}}\, \mathcal{L}_{\text{att}} + \lambda_{\text{fid}}\, \mathcal{L}_{\text{fid}}. \tag{2}$$

$\mathcal{L}_{\text{ans}}$ is the answer cross-entropy. $\mathcal{L}_{\text{att}}$ is the mean over fixation heads of the negative log of each head's share inside the annotated box of the organ the question asks about, the correct option for comparison questions, the quantity that discovery measures. $\mathcal{L}_{\text{fid}}$ is a KL term to the frozen model on prompts that name no region, which keeps general behaviour intact. Organ boxes are needed on training images only, and heads ranked on training images alone give the same gain (Appendix D.5). At test time no region and no mask are given, and Appendix B.6 lists r, α, the loss weights and the schedule.

4 Experimental Setup

Models. The roster has 25 instruction-tuned VLMs from 3 families, Qwen2.5-VL (Bai et al., 2025a), Qwen3-VL (Bai et al., 2025b) and Gemma3 (Team et al., 2025), from 2B to 73B parameters, and Qwen-VL denotes both Qwen generations. 14 medical models are tunes of a base that the roster also contains, 4 supervised fine-tunes and 10 reinforcement-learning or grounding tunes, so head indices match one to one. Discovery and ACG region sources run on all 25 models, steering and most read-out tests on a core set of 17 models that spans all three families and every size, MedGaze on 14 models, bases and tunes, and report generation on 10 of these, as Appendix Figure 7 maps. All models run in `bfloat16` with eager attention.

Data. Discovery uses montages from VQA-RAD (Lau et al., 2018), with PathVQA (He et al., 2020) and SLAKE (Liu et al., 2021) montages as modality checks. Grounding uses radiologist boxes on MS-CXR (Boecking et al., 2022), VinDr-CXR (Nguyen et al., 2022), Chest ImaGenome (Wu et al., 2021) and SLAKE. Organ VQA and MedGaze use SLAKE, split by image so that test images are never seen in training, and report generation uses 199 held-out Chest ImaGenome studies per model, whose findings a rule-based labeller with negation handling scores against the study's Chest ImaGenome findings as clinical finding F1. The *region tour* prompts one sentence per chest region, with the heads steered to its anatomy box or left alone. Credentialed data are processed locally under their data use agreements, and Appendix C gives all settings. Every study shown as a qualitative example was reviewed by an expert radiologist.

Baselines. Random heads and steering all heads test head specificity. ACG's region sources are compared with a 3×3 grid, Segment Anything and the annotated box, and the read-out maps with the mean over all heads, rollout and gradient attribution. MedGaze is compared with an answer-only adapter and the decoders VCD, PAI and CRG (Leng et al., 2024; Liu et al., 2024; Wan et al., 2024).

Statistics. One model is one observation for every cross-model claim, tested with paired Wilcoxon tests across models or exact McNemar tests on paired answers where stated, with Cohen's d for per-finding effects and Holm correction within each family of metrics. *Chance* is the value of attention with no information about the named region, $1/N$ over N regions, the organ's share of image tokens along generation, one over the candidate regions for an ACG hit, and two random head sets' overlap.

5 Results and Discussion

The subsections test in turn whether medical VLMs have fixation heads and keep them through tuning, whether the heads are steerable, whether their attention can be read as provenance, whether writing the region gives provenance, whether the heads can learn to look with no region, and whether both writes carry over to reports.

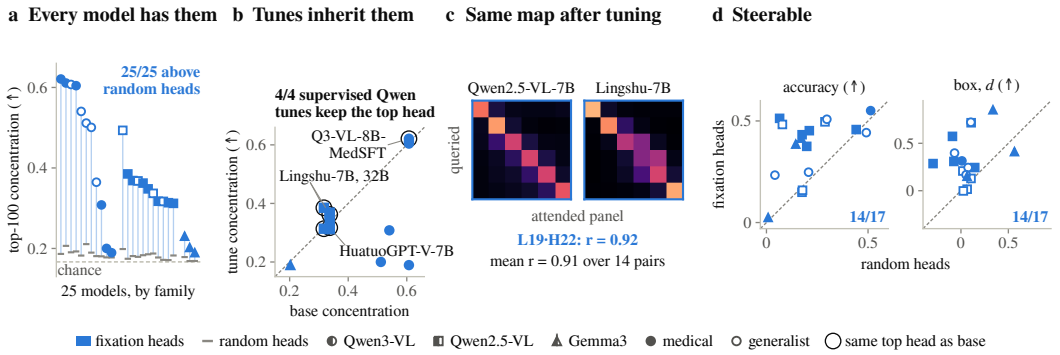


Figure 3: **Every model has fixation heads, supervised tunes inherit them, and they are steerable.** (a) Top-100 against random heads; (b) each tune against its base; (c) one head's fixation matrix in base and tune; (d) steering, random heads (x) against fixation heads (y).

5.1 Fixation heads exist and are inherited

Every model has fixation heads. On montages, the top-100 heads reach a concentration of 0.189 to 0.621 on the queried panel, against a chance level of 0.167 and 0.169 to 0.211 for 100 random heads, so fixation heads lead in all 25 models, as Figure 3a shows. Scored on held-out PathVQA montages that never informed their selection, the same heads lead in all 17 of 17 models ($p = 2 \times 10^{-5}$).

Supervised medical tunes inherit them. All 4 of 4 supervised Qwen tunes paired with their base in the roster keep the base's top head, against 1 of 10 other tunes (Figure 3b). Across all 14 pairs, a tune shares a median 0.80 of its base's top-10 fixation heads against at most 0.09 by chance. Fine-tuning preserves the fixation matrices themselves, as the entries of the top head's matrix in base and tune correlate at $r = 0.91$ over 14 pairs, against 0.80 for the same random head and 0.42 for mismatched heads (Wilcoxon $p = 1.2 \times 10^{-4}$), as Figure 3c shows for one pair, head L19·H22 of `Qwen2.5-VL-7B` and `Lingshu-7B` ($r = 0.92$).

5.2 The heads are steerable

Steering redirects the description. Steered to one montage panel, the fixation heads make a fixed judge, an unsteered `Qwen3-VL-8B` that is never told the target, pick that panel in 37.2% of cases against 21.7% for random heads and 16.7% chance, and they lead in 14 of 17 models ($p = 0.0017$), which are the points above the diagonal in Figure 3d, so the effect is specific to the discovered heads; with the first token steered as well, they reach 39.6% against 19.8% on 15 models (Appendix B.3). Biasing *all* heads instead collapses accuracy to 1.3%, with degenerate text in 88 to 100% of outputs, so the write needs a head selection.

Steering works within one image. For each MS-CXR finding, the phrase's log-probability under steering to the true box minus that to an equal-area control box ($b = 10^4$) cancels image content and phrase length, and d is Cohen's d over findings. Fixation heads give $d = +0.36$ against $+0.08$ for random heads, ahead in 14 of 17 models ($p = 0.0079$; Figure 3d), and steering to SLAKE boxes repeats the effect ($d = +0.53$ against $+0.13$, 15 of 17 models, $p = 0.00021$).

5.3 The heads are not readable

Five tests read the heads' attention as evidence of what the model used, and all five fail. A map is *readable* as provenance when the region a claim came from holds most of the heads' share, so that the map points a reader at it.

Attention barely follows the question. Discovery shows that the heads follow a positional query across stitched images, whereas reading asks whether naming an organ in one scan moves them. Fixation-head maps for different questions about the *same* image correlate at 0.916, against 0.417 for one question on *different* images, in 17 of 17 models ($p = 2 \times 10^{-5}$). The heads do prefer a named organ, with a concentration of 0.423 against 0.333 chance on segmented SLAKE anatomy

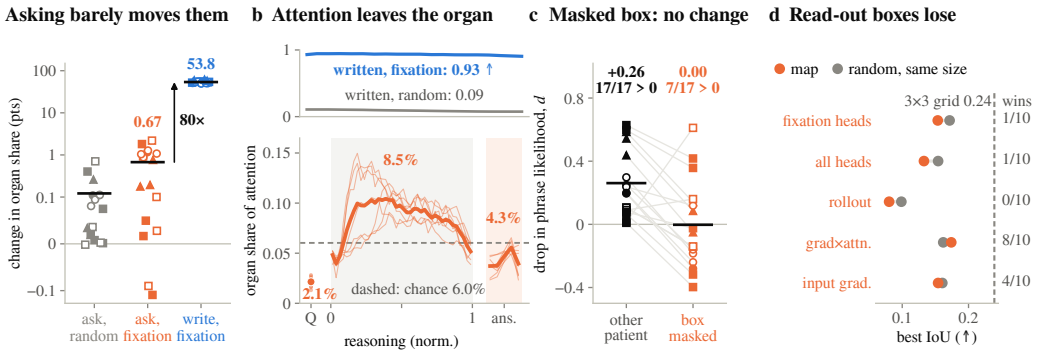


Figure 4: **Attention written into the heads holds, whereas attention read from them does not show what the model used.** (a) Symmetric log axis; (b) written heads inset; (c) masking the radiologist's box leaves the phrase unchanged on average; (d) box proposals.

(1.27×), yet the share this preference buys is small. On SLAKE scans with several organ masks, asking about an organ instead of another raises the heads' share on it by 0.67 ± 0.70 points from a base of 1.9%, positive in 15 of 17 models (Holm $p = 0.0035$), whereas writing the organ at $b = 4$ raises it by 53.8 ± 3.2 points (mean $\pm$ sd over models, Figure 4a).

Attention leaves the organ as the model answers. The heads' share of image attention on the organ, followed token by token while 7 models prompted to reason before answering work through a SLAKE organ question, peaks during reasoning, as Figure 4b shows against a dashed line at the organ's 6.0% share of the image, whereas steered heads hold a share of 0.93 throughout (inset). The share is 2.1% at the question, rises to 8.5% while the model reasons, and falls to 4.3% on the answer, and reasoning exceeds the answer in all 7 models ($p = 0.016$, the minimum for seven) and in 295 of 336 questions. A map taken at the answer token therefore misses the organ the same heads attended during reasoning, so where a map is read changes the evidence it shows.

Masking the region leaves the phrase unchanged. Masking the box a radiologist drew leaves the likelihood of that radiologist's phrase unchanged (mean $d \approx 0$ over 17 models), whereas a different patient's image lowers it ($d = +0.26$) in all 17 models, as Figure 4c shows, and SLAKE and VinDr-CXR agree. If the attended region were the evidence the unsteered model used, removing its pixels would lower the phrase, so its attention marks a target rather than a source, whereas once the region is written into the heads its content drives the sentence (§5.4).

Attention proposes no better regions than random boxes. Over 10 models and 400 MS-CXR findings, boxes proposed from the fixation heads' attention reach a mean best IoU with the radiologist's box of 0.154 against 0.173 for size-matched random boxes, as Figure 4d shows, where wins count the models above random before Holm correction, and a single held-out proposer loses at all 3 thresholds (Appendix D.3). The mean over all heads, attention rollout (Abnar & Zuidema, 2020), gradient×attention (Chefer et al., 2021) and input-gradient saliency do not beat their own size-matched random boxes after Holm correction, the closest being gradient×attention (+7%, 8 of 10 models, $p = 0.15$), and a fixed 3×3 grid, whose nine cells give it more proposals than any map, beats every map (IoU 0.239).

Attention does not flag errors. The heads' focus detects hallucination no better than random heads and reranks sampled answers no better than log-probability (Appendix D.3).

5.4 Writing the region with anatomy-conditioned generation

The heads look where they are sent, and the sentence follows. Steered to an organ mask, the fixation heads place a share of 0.90 to 0.99 inside it, with a median of 0.98 over 151 steered regions in 11 models, and 64% of steered sentences name the organ they were steered to (14 models), as in the example of Figure 5a. Blanking the steered region while the heads stay on it, in 8 models with 454 steered regions each, lowers how often the sentence names the organ from 52% to 22%, against

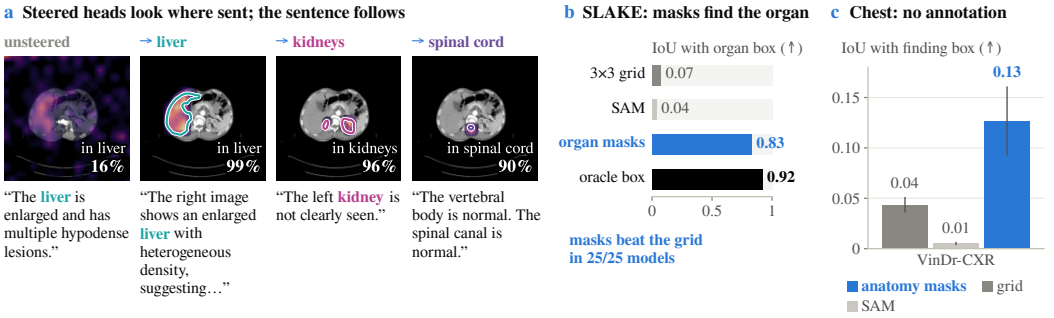


Figure 5: **Steered fixation heads look where they are sent, and organ masks select the region without annotation.** (a) `Lingshu-7B` steered to three organ masks of one SLAKE CT scan; IoU by region source (b) with the organ box on SLAKE and (c) with the finding box on VinDr-CXR.

Table 1: **MedGaze outperforms training-free decoders with no region at test, and its attention term moves the heads at the accuracy of the answer loss alone.** Held-out SLAKE organ questions: (A) 5 models, (B) all 14, (C) a second image split; bold: best per block; n/a: not recorded for decoders, set by the mask for steering; parameters: mean per block.

Method	region at test	trainable params	accuracy (%) ↑	organ-choice acc. (%) ↑	yes/no questions bal. acc. (%) ↑	yes/no questions MCC ↑	attention in organ ↑	heads peaking in organ ↑
A. Same 5 models, 210 held-out questions each (mean±sd over models)								
Frozen	none	0	70.5±9.3	65.7±9.3	74.0±11.9	0.45±0.06	0.07±0.03	0.07±0.04
VCD (Leng et al., 2024)	none	0	71.2±8.4	65.3±10.6	74.0±6.4	0.42±0.08	n/a	
PAI (Liu et al., 2024)	none	0	55.6±22.3	55.3±31.3	58.6±25.9	0.32±0.06	n/a	
CRG (Wan et al., 2024)	mask	0	69.1±8.5	67.3±9.2	66.9±6.8	0.34±0.12	n/a	
Fixation steering, b=4	mask	0	72.9±8.2	70.3±9.7	76.4±7.2	0.46±0.06	n/a	
Answer-only adapter	none	0.79M	**84.7±2.6**	72.0±4.8	**81.2±3.1**	**0.61±0.04**	0.07±0.03	0.06±0.04
MedGaze (ours)	none	0.79M	83.9±2.8	**72.3±6.1**	80.5±2.9	0.58±0.04	**0.31±0.10**	**0.41±0.10**
B. All 14 models with an adapter (mean±sd over models)								
Frozen	none	0	75.1±9.5	73.1±11.0	75.8±12.1	0.43±0.18	0.09±0.06	0.10±0.08
MedGaze (ours)	none	0.73M	**84.8±3.4**	**73.7±6.5**	**81.2±5.0**	**0.60±0.10**	**0.33±0.11**	**0.45±0.12**
models improved			12/14	8/14	8/14	13/14	14/14	14/14
C. Loss terms, retrained on a second image split (mean over 4 models; change against frozen)								
Frozen	none	0	73.5	66.4	75.2	0.39	0.09	0.10
Answer loss only	none	0.67M	**82.1 (+8.6)**	73.0	**77.5**	**0.49**	0.10 (1.1×)	0.11
+ attention term: MedGaze (ours)	none	0.67M	81.8 (+8.3)	**73.8**	76.0	0.46	0.34 (3.7×)	0.46
Reliance variant, 200 steps	none	0.67M	76.7 (+3.2)	69.7	76.0	0.42	**0.38 (4.1×)**	**0.51**

50% for a random region, and in the steered region tour it costs 9.2 F1 points more than the same blanking costs the unsteered tour ($p = 0.012$, 8/9 models, Appendix D.4).

Label-free regions localise the finding. On chest radiographs the regions come from a segmenter run on the test image, and its masks select the finding's region at 1.8× chance on MS-CXR and 2.0× on VinDr-CXR, with tighter regions than the grid on VinDr-CXR (IoU 0.127 against 0.044, 24 of 25 models), as Figure 5c shows, with at most 5.0% of steered outputs degenerate. On SLAKE, whose organ boxes derive from the same masks, the region chosen from organ masks reaches IoU 0.830, near the oracle box (0.919, below 1 because every region snaps to the token grid) and far above a 3×3 grid (0.075), which is the tightness ACG reaches when the masks are exact (Figure 5b).

5.5 Learning to look with MedGaze

MedGaze outperforms training-free decoders with no region. On held-out SLAKE organ questions under the same prompt and decoder, MedGaze is more accurate than VCD, PAI, CRG and fixation steering in all 5 models compared, on both accuracy and the Matthews correlation coefficient (MCC), although CRG and steering see the organ mask (Table 1, block A). An answer-only adapter trained on the same split is as accurate but leaves the heads' share at the frozen level, so the attention term is what moves the heads.

MedGaze moves the heads onto the anatomy. MedGaze raises the fixation heads' share inside the answer organ from 0.09 to 0.33, and the share of heads that peak inside it from 0.10 to 0.45, in all 14 of 14 models, as block B of Table 1 and Figure 6a show. Accuracy rises from 75.1 to 84.8% (+9.7

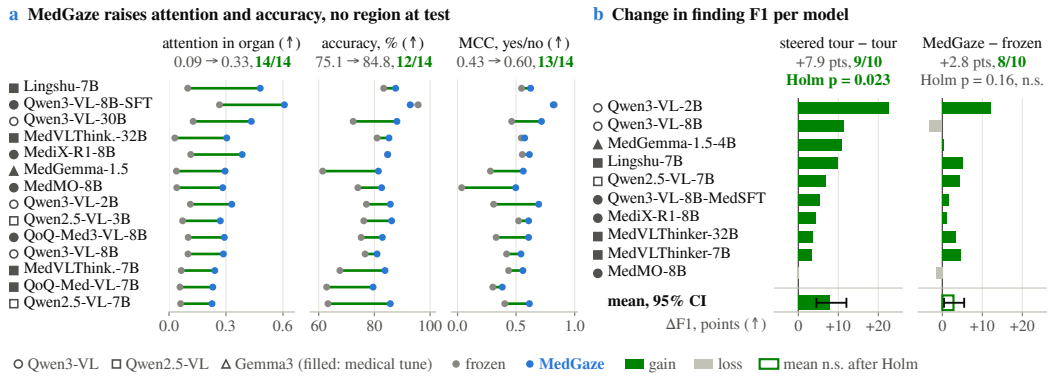


Figure 6: **MedGaze moves attention onto the organ and raises accuracy with no region, and writing the region adds findings.** (a) Frozen (grey) against MedGaze (blue) on SLAKE, names as in Appendix Table 9; (b) F1 change per model; hollow: n.s. after Holm.

points) in 12 of 14 models, and MCC, which a constant answer cannot inflate, rises from 0.43 to 0.60 in 13 of 14 models. Appendix Table 9 adds the yes rate and visual reliance.

5.6 Writing the region into reports

Steering adds findings. Steering raises the tour's clinical finding F1 ($\times 100$) from 19.7 to 27.6 ($+7.85$) in 9 of 10 models (Holm $p = 0.023$), as Figure 6b shows, with recall up $+10.97$ while the tour grows only from 53 to 57 words, so writing the region changes which findings a report mentions. MedGaze alone, with no region, lifts mean F1 from 24.4 to 27.2 points, higher in 8 of 10 models ($p = 0.049$, Holm $p = 0.16$), and BLEU and ROUGE do not move.

5.7 Ablations and statistical analysis

Which heads and which loss. Random heads trained with the same recipe end with less attention on the organ in 14 of 14 models. In block C of Table 1, retrained on a second image split, the answer loss alone reaches the same accuracy but leaves attention unchanged ($1.1\times$), whereas the attention term raises it $3.7\times$ at the accuracy of the answer loss alone. The *reliance variant* adds a counterfactual hinge, requiring the answer's log-probability to fall when the organ is mean-filled, and makes the answer depend on the organ in 4 of 4 models (Appendix Table 10).

Robustness. In the two models with a clear default effect, all 14 alternative steering settings stay significant, and every MedGaze change beats seed noise (Appendix D.5).

6 Conclusion

This paper introduces fixation heads, the heads that route a medical VLM's attention to a named region, as the unit of provenance. They exist in every model tested, and supervised medical tunes keep them from their base. Steered to a region, they redirect what the model describes more often than random heads in 14 of 17 models, yet their attention barely follows the question and does not mark what the model used, because a head's attention is an input to the computation, not a record of it. Anatomy-conditioned generation turns this write access into provenance, and blanking the written region costs the steered tour 9.2 more F1 points than it costs the unsteered one. MedGaze learns the same access with no region at test, raising the heads' attention on the organ from 0.09 to 0.33 and accuracy from 75.1 to 84.8%. Its attention loss is supervised with SLAKE organ boxes, and supervising it with chest anatomy is the natural next step. We encourage writing provenance during decoding instead of reading it from attention maps.

AI USE STATEMENT

We used generative AI tools in two ways. They assisted with literature exploration, and with polishing the writing of this paper. All AI-assisted content was reviewed and verified by the authors. The authors take full responsibility for the content of this paper.

ETHICS STATEMENT

MS-CXR, Chest ImaGenome and MIMIC-CXR are credentialed data used under their data use agreements. VinDr-CXR, SLAKE, VQA-RAD and PathVQA are used under their licences. All credentialed data were processed locally and never sent to an external service. No reference report text is reproduced, and the credentialed radiographs in the figures are shown as their data use agreements permit. The methods are research tools for studying model internals. They are not validated for clinical decisions, and a steered or adapted model can still state wrong findings.

REPRODUCIBILITY STATEMENT

The appendix gives every method detail needed to reimplement the work (Appendix B). It covers the discovery score, the region protocols, steering, anatomy-conditioned generation, both learned stages and the evaluation protocols. Datasets, model checkpoints and statistics are listed in Appendix C. Every number in the paper is emitted by aggregation scripts from the run files. Upon publication we will release the code, a Python package that implements discovery, steering and MedGaze, and the trained adapters (Appendix E).

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4190–4197, 2020.

Estelle Aflalo, Gabriela Ben Melech Stan, Tiep Le, Man Luo, Shachar Rosenman, Sayak Paul, Shao-Yen Tseng, and Vasudev Lal. Fivl: A framework for improved vision-language alignment through the lens of training, evaluation and explainability. *arXiv preprint arXiv:2412.14672*, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025b.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pp. 1–21. Springer, 2022.

Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 387–396. IEEE, 2021.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, et al. Towards injecting medical visual knowledge into multimodal LLMs at scale. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 7346–7370, 2024.

Joseph Paul Cohen, Joseph D Viviano, Paul Bertin, Paul Morrison, Parsa Torabian, Matteo Guarrera, Matthew P Lungren, Akshay Chaudhari, Rupert Brooks, Mohammad Hashir, et al. TorchXRayVision: A library of chest X-ray datasets and models. In *International Conference on Medical Imaging with Deep Learning*, pp. 231–249. PMLR, 2022.

Parsa Esmaeilkhani and Longin Jan Latecki. Direct visual grounding by directing attention of visual tokens. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5787–5797. IEEE, 2026.

Rohit Gandikota and David Bau. Gaze heads: How vlms look at what they describe. *arXiv preprint arXiv:2606.14703*, 2026.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13418–13427. IEEE, 2024.

Xiaoke Huang, Juncheng Wu, Hui Liu, Xianfeng Tang, and Yuyin Zhou. Medvlthinker: Simple baselines for multimodal medical reasoning. *arXiv preprint arXiv:2508.02669*, 2025.

Sarthak Jain and Byron C Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, 2019.

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.

Sonia Joseph, Praneet Suresh, Lorenz Hufe, Edward Stevinson, Robert Graham, Yash Vadi, Danilo Bzdok, Sebastian Lapuschkin, Lee Sharkey, and Blake Aaron Richards. Prisma: An open source toolkit for mechanistic interpretability in vision and video. *arXiv preprint arXiv:2504.19475*, 2025.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. Your large vision-language model only needs a few attention heads for visual grounding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9339–9350. IEEE, 2025a.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *International Conference on Learning Representations*, volume 2025, pp. 87676–87703, 2025b.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

Zixuan Lan, Luzhe Sun, Matthew R Walter, and Jiawei Zhou. Seeing without looking: Do vision-language benchmarks really test vision? *arXiv preprint arXiv:2605.22903*, 2026.

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.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882. IEEE, 2024.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in neural information processing systems*, 36:41451–41530, 2023.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pp. 1650–1654. IEEE, 2021.

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

Zhining Liu, Ziyi Chen, Hui Liu, Chen Luo, Xianfeng Tang, Suhang Wang, Jingying Zeng, Zhenwei Dai, Zhan Shi, Tianxin Wei, et al. Seeing but not believing: Probing the disconnect between visual attention and answer correctness in vlms. In *International Conference on Learning Representations*, volume 2026, pp. 72849–72876, 2026.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2034–2044. IEEE, 2025.

Mahshad Lotfinia, Sebastian Ziegelmayer, Lisa Adams, Daniel Truhn, Andreas Maier, and Soroosh Tayebi Arasteh. Vision-language models for chest radiography do not always need the image. *arXiv preprint arXiv:2606.17710*, 2026.

Tiange Luo, Lajanugen Logeswaran, Jaekyeom Kim, Justin Johnson, and Honglak Lee. Selective lora for visual tokens and attention heads. *arXiv preprint arXiv:2512.19219*, 2025.

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

Sahal Shaji Mullappilly, Mohammed Irfan Kurpath, Omair Mohamed, Mohamed Zidan, Fahad Khan, Salman Khan, Rao Anwer, and Hisham Cholakkal. Medix-r1: Open ended medical reinforcement learning. *arXiv preprint arXiv:2602.23363*, 2026.

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.

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 337–347. Springer, 2025.

Filippo Ruffini, Marco Salmé, Rosa Sicilia, Valerio Guarrasi, and Paolo Soda. Shovir: A benchmark for evaluating vision shortcut learning in radiology report generation. *arXiv preprint arXiv:2606.30201*, 2026.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. MedGemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

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

Ramprasaath R Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh. Taking a hint: Leveraging explanations to make vision and language models more grounded. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2591–2600, 2019.

Sofia Serrano and Noah A. Smith. Is attention interpretable? In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2931–2951, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1282. URL https://aclanthology.org/P19-1282/.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.

Jongmin Shin, Ka Young Kim, Eunki Cho, Seong Tae Kim, and Namkee Oh. Surgcheck: Do vision–language models really look at images in surgical vqa? *International Journal of Computer Assisted Radiology and Surgery*, pp. 1–9, 2026.

Pratham Singla, Shivank Garg, Vihan Singh, and Paras Chopra. Do vision-language models see or guess? measuring and reducing textual-prior reliance with a phrasing-controlled benchmark. *arXiv preprint arXiv:2606.10400*, 2026.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

David Wan, Jaemin Cho, Elias Stengel-Eskin, and Mohit Bansal. Contrastive region guidance: Improving grounding in vision-language models without training. In *European Conference on Computer Vision*, pp. 198–215. Springer, 2024.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 11–20, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1002. URL https://aclanthology.org/D19-1002/.

Joy T Wu, Nkechinyere N Agu, Ismini Lourentzou, Arjun Sharma, Joseph A Paguio, Jasper S Yao, Edward C Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, et al. Chest imagenome dataset for clinical reasoning. *arXiv preprint arXiv:2108.00316*, 2021.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. In *International Conference on Learning Representations*, volume 2025, pp. 62143–62156, 2025.

Huihui Xu, Yuanpeng Nie, Hualiang Wang, Ying Chen, Wei Li, Junzhi Ning, Lihao Liu, Hongqiu Wang, Lei Zhu, Jiyao Liu, et al. Medground-r1: Advancing medical image grounding via spatial-semantic rewarded group relative policy optimization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 391–401. Springer, 2025a.

Tianrun Xu, Haoda Jing, Ye Li, Yuquan Wei, Jun Feng, Guanyu Chen, Haichuan Gao, Tianren Zhang, and Feng Chen. Defacto: Counterfactual thinking with images for enforcing evidence-grounded and faithful reasoning. *arXiv preprint arXiv:2509.20912*, 2025b.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: Generalist foundation model for unified multimodal medical understanding and reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Zhongzhi Yu, Zheng Wang, Yonggan Fu, Huihong Shi, Khalid Shaikh, and Yingyan Celine Lin. Unveiling and harnessing hidden attention sinks: Enhancing large language models without training through attention calibration. *arXiv preprint arXiv:2406.15765*, 2024.

Jianrui Zhang, Mu Cai, Tengyang Xie, and Yong Jae Lee. Countercurate: Enhancing physical and semantic visio-linguistic compositional reasoning via counterfactual examples. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15481–15495, 2024a.

Qingru Zhang, Chandan Singh, Liyuan Liu, Xiaodong Liu, Bin Yu, Jianfeng Gao, and Tuo Zhao. Tell your model where to attend: Post-hoc attention steering for llms. In *International Conference on Learning Representations*, volume 2024, pp. 42411–42430, 2024b.

Zifan Zheng, Yezhaohui Wang, Yuxin Huang, Shichao Song, Mingchuan Yang, Bo Tang, Feiyu Xiong, and Zhiyu Li. Attention heads of large language models: A survey. *arXiv preprint arXiv:2409.03752*, 2024.

APPENDIX OVERVIEW

Appendix A gives the background and notation, Appendix B the full method and Appendix C the experimental details. Appendix D extends each result of Section 5 in order, Appendix E lists what is needed to reproduce it, and Appendix F collects qualitative examples that an expert radiologist reviewed.

A BACKGROUND

Why current methods do not attach provenance to each claim. Three families of tools address provenance, and none attaches a region to each claim. Behavioural audits perturb the input and show that a model under-uses the image, but they do not locate the component that routes regions or say how to change it. Attention and attribution maps are drawn after the fact and read as a record of what the model used, yet this reading fails in medical VLMs for the fixation heads, rollout and gradient attribution alike, as Appendix D.3 shows. Attention steering and supervision bias all heads, retrain broadly or rely on natural-image boxes, so they change outputs without tying a region to each sentence. Provenance requires both parts at once, a component that routes regions and a procedure that uses it to tie every sentence to the anatomy it describes.

Steering, supervising and adapting attention. Changing where a model looks is the other half of the problem. Visual contrastive decoding (VCD) contrasts the output against a noised image (Leng et al., 2024), PAI amplifies image attention (Liu et al., 2024), and contrastive region guidance (CRG) contrasts the image against one with the region blacked out (Wan et al., 2024), with related decoders in Huang et al. (2024) and Yu et al. (2024). Training-time methods align attention to human or box maps (Selvaraju et al., 2019; Esmaeilkhani & Latecki, 2026; Aflalo et al., 2024), reward grounding (Xu et al., 2025a; Shen et al., 2025; Liu et al., 2025), adapt selected heads (Luo et al., 2025) or use counterfactuals (Zhang et al., 2024a; Xu et al., 2025b). These methods act on all heads, retrain broadly, or use natural-image boxes, whereas our work updates only behaviourally discovered heads, only at decoding, with annotated organ boxes as supervision.

Task and notation. A vision-language model (VLM) receives an image x and a prompt q and generates an answer y token by token. The image becomes a set $\mathcal{I}$ of image-token positions, and a region R of the image becomes the set $T(R) \subseteq \mathcal{I}$ of tokens that the token-region rule of Appendix B.2 assigns to it. The language model has L layers of H query heads, and head h at position t has post-softmax attention $\alpha^h_{t,j}$ over earlier positions j. The final prompt token t^* is the position where the model has read the question and is about to answer. The attention mass of head h on region R is $m^h_t(R) = \sum_{j \in T(R)} \alpha^h_{t,j}$.

Fixation heads. Take N disjoint regions $R_1, \ldots, R_N$ of one input and a query q_i that names R_i. The fixation matrix of head h has entries $A_h(i,j) = m^h_{t^*}(R_j \mid q_i)$, and, following Gandikota & Bau (2026), the fixation score is its mean diagonal,

$$s_h = \frac{1}{N} \sum_{i=1}^{N} A_h(i,i). \tag{3}$$

The mass is raw and never renormalised across regions, so a head that ignores the image scores zero. The 100 heads with the highest score form the fixation set $\mathcal{H}$, and 100 heads drawn with a fixed seed from outside $\mathcal{H}$ are the random heads. The concentration of a head set is the mean over its heads of $\frac{1}{N} \sum_i A_h(i,i) / \sum_j A_h(i,j)$, with chance level $1/N$.

Reading and writing. To *read* the heads is to treat their attention as evidence of what the model used, as an attention map, a box derived from it, a hallucination score or a reranking score does. To *write* a region into the heads is to change their attention logits so that they attend to it, as steering, anatomy-conditioned generation (ACG), calibrated fixation steering and the MedGaze adapter do. MedGaze writes through training, so it needs no region at test time.

Six questions. Appendix D follows the main text in order, testing whether medical VLMs have fixation heads and keep them through fine-tuning, whether writing a region into the heads changes what the model says, whether their attention shows what the model used, whether writing the region at inference gives provenance, whether a model can learn to look at the anatomy with no region at test, and whether both ways of writing carry over to free-text reports.

B Extended Method

The procedures follow the order of Section 3, and the evaluation protocols come last.

B.1 Discovery score

Montage protocol. Six unrelated images of one corpus are stitched into a horizontal strip, 448 px high with an 8 px gap, and panel edges snap to the token grid so that no token straddles two panels. The query names a panel by position only ("what do you see in the k-th image?"), so the answer region is known by construction and no anatomical prior can help. Each model sees 200 montages with six queries each, 1,200 prefill passes in total as Appendix C lists, on VQA-RAD for discovery and on PathVQA and SLAKE montages for transfer. One VQA-RAD montage queried for its third panel, with the share of attention that the fixation heads and 100 random heads outside $\mathcal{H}$ put on the queried panel, forms the top row of Figure 8.

The same score, Equation 3, ranks heads on any N named regions of one image, and every run uses eager attention so that attention is materialised.

Which ranking each experiment uses. Discovery, montage steering and the transfer tests use the montage ranking. Experiments on an annotated corpus (phrase steering, ACG, MedGaze and report generation) use the ranking by grounding gap on that corpus's own boxes where one exists, and the montage ranking otherwise. Ranking the SLAKE heads on training images only keeps 0.88 to 0.99 of the top 100 heads.

B.2 Region protocols and the token-region rule

Three protocols. Besides the positional montage, radiologist boxes on MS-CXR, VinDr-CXR, Chest ImaGenome and SLAKE score a head by its enrichment and grounding gap, defined next, and heads ranked on the boxes themselves are called *box-ranked*.

The *enrichment* of head h in box B is its share inside the box divided by the box's share of the image, $e_h(B) = \frac{m^h_{t*}(B)/m^h_{t*}(\mathcal{I})}{|B|/|\mathcal{I}|}$, so that uniform attention scores 1, and the *grounding gap* is $g_h = e_h(B) - e_h(B')$ for the matched control box B', averaged over findings. Box-ranked heads are the 100 heads with the largest gap.

Segmented anatomy applies the $N \times N$ protocol to one unmodified image, with named structures from a chest anatomy segmenter (Cohen et al., 2022) or SLAKE's organ masks, and its control is a positional query that differs only in wording ("the third region from the left").

Control boxes. Three controls are scored on the same forward pass, (i) a *matched* box of the same size at a random location with IoU at most 0.05 with the annotated box, (ii) a *shifted* box, the annotated box moved about 1.25 box widths and clipped, again with IoU at most 0.05, so that it stays in the same neighbourhood, and (iii) a *wide* box of twice the area on the same centre.

Token-region rule. A token joins a region when at least 25% of its patch lies inside the mask, computed per pixel, and overlaps go to the region with the larger cover in discovery, whereas steering, ACG and MedGaze map each region on its own, so overlapping regions share tokens. If no token reaches the threshold, the threshold is halved and then the single best token is kept, and the recorded token count lets every analysis condition on it. Masks are used whenever they exist, because a bounding box would leak an irregular organ onto its neighbours.

SLAKE masks. SLAKE stores organ identity as unlabelled grey levels, and matching each level's extent to the detection boxes over the corpus recovers the legend. Each level is validated per image (precision at least 0.6 inside a 10%-dilated box, recall at least 0.2), and a level that fails falls back to its rectangle.

B.3 Steering

The intervention. Steering adds $+b$ to the pre-softmax logits of the image tokens in $T(R)$ and $-b$ to those in $\mathcal{I} \setminus T(R)$ for each head in $\mathcal{H}$, while heads outside $\mathcal{H}$, including those in the same layer,

Algorithm 1 Calibrated fixation steering

Require: fixation heads $\mathcal{H}$, 48 calibration items $\mathcal{D}$

1: $\theta_h \leftarrow \sigma^{-1}(b_0/g_{\max})$ for all $h \in \mathcal{H}$ ▷ start at b_0
2: **for** $t = 1, \ldots, 40$ **do**
3: $B \leftarrow$ 8 items $(x, R, q, y^+, \{y^-\})$ from $\mathcal{D}$
4: prefill (x, q) but its last token ▷ no gradient
5: $g_h \leftarrow g_{\max}\,\sigma(\theta_h)$ for all $h \in \mathcal{H}$
6: $z \leftarrow$ logits of the last token, bias $\pm g_h$ on R
7: $\ell \leftarrow \mathrm{CE}(z_{\{y^+\}\cup\{y^-\}}, y^+) + \lambda\|g - b_0\|^2$
8: $\theta \leftarrow \mathrm{Adam}(\theta, \nabla_\theta \ell)$ ▷ lr 0.1, $\lambda = 10^{-3}$
9: **end for**
10: **return** $(\mathcal{H}, g)$ ▷ a few kilobytes

keep the original logits. Under grouped-query attention the bias is indexed by query head after key and value expansion, so it broadcasts correctly.

Operating points. At $b = 10^4$ the bias saturates in `bfloat16` and acts as a hard gate, which scores a fixed phrase and runs the montage test. Free generation uses $b = 4$, the largest value at which every model still writes clean text, as Table 4 shows.

Split prefill. Because steering hooks act on decoding steps only while the first generated token is predicted from the prefill row, the decoder prefills all but the last prompt token, runs the last prompt token as one steered step, and decodes greedily with the checkpoint's own repetition penalty, in the steered and unsteered conditions alike. The montage test generates with plain greedy decoding. With the steered first token on 15 models, the fixation heads reach 39.6% against 19.8% for random heads and lead in 12 of 15, so the split prefill only strengthens the effect. ACG scores the reference phrase with the steered first token, so its hit rate and IoU do not depend on the decoding path.

Phrase scoring. Region-of-interest steering scores the log-probability of the radiologist's phrase with the heads steered to its box against the same heads steered to an equal-area control box, and the effect is Cohen's d over findings for fixation heads and for random heads.

B.4 Anatomy-conditioned generation

Region sources. Four region sources run on identical images, models, heads and bias, so that their comparison is the ablation, (i) a 3×3 grid with no anatomy in it, (ii) the annotated box as an oracle, (iii) SLAKE's organ masks, or on chest radiographs the 14 structures of the anatomy segmenter, and (iv) Segment Anything (Kirillov et al., 2023) in automatic mode, class-agnostic, keeping the top 40 segments by predicted quality.

Procedure and selection rule. ACG steers the heads to each region in turn at $b = 4$ and generates one response per region with a fixed prompt, and at deployment every sentence is kept with its mask. For evaluation, the selected region is the one under whose steering the reference phrase, scored directly after the prompt, is most likely. An optional refinement splits the selected mask into four compact parts by k-means on its pixels and repeats the step.

Metrics. A *hit* means that the selected region contains the box, and chance is the mean of one over the number of regions, which differs by source. A *lenient hit* requires only that the selected region overlaps the box. IoU compares the selected region with the box, and the *ceiling* is the best IoU that any region of that source could reach. Text quality is measured by degeneracy, repeated 4-grams and region specificity, which Appendix B.8 defines.

B.5 Calibrated fixation steering

Calibrated fixation steering learns one scalar gain per fixation head, with the model frozen and the region given. The gain is $g_h = g_{\max}\,\sigma(\theta_h)$ with $g_{\max} = 12$, initialised at the fixed bias $b_0 = 4$, so that at step zero the method *is* fixed-bias steering.

Algorithm 1 lists the fit. Every gain starts at the fixed bias (line 1), so at step zero the method is fixed-bias steering. Each of 40 steps prefills a batch of 8 items without gradient except for the last prompt token (lines 3 and 4), runs that token with each head steered by its own gain (lines 5 and

6), and scores the correct candidate against the others with a pull towards b_0 (line 7). Only the gain parameters θ change (line 8), and $\mathcal{H}$ comes from the SLAKE box ranking, or from the montage ranking where none exists. The candidates are "Yes" and "No", or the named organ against the other organs in the image. Calibration uses a quarter of SLAKE's images, split by image, and evaluation runs on the rest with the learned gains as one more condition, so the comparison with the fixed bias is paired on identical questions. A fit takes at most 11 minutes on one GPU.

B.6 The MedGaze adapter

Architecture. In each layer that holds a fixation head, the query projection becomes

$$q = W_q x + \tfrac{\alpha}{r}\, M\, B\, A\, x, \qquad r = 8,\ \alpha = 16, \tag{4}$$

where W_q is frozen, A is a rank-8 down-projection, B is zero-initialised, and M scatters BA into the query rows of the fixation heads only (Hu et al., 2021). Keys, values and every other head stay bit-identical, and the adapter acts at decoding only, so the cached prompt is the frozen model's. The adapter has 0.45M to 1.00M parameters.

Losses. At the answer step, the loss of Section 3.5 uses $\lambda_{\text{att}} = 0.25$ and $\lambda_{\text{fid}} = 0.1$,

$$\mathcal{L} = \mathcal{L}_{\text{ans}} + 0.25\, \mathcal{L}_{\text{att}} + 0.1\, \mathcal{L}_{\text{fid}}. \tag{5}$$

$\mathcal{L}_{\text{ans}}$ is the teacher-forced cross-entropy on the answer. $\mathcal{L}_{\text{att}}$ is $-\log$ of each head's share of the fixation heads' image attention inside the annotated box of the organ the question asks about (the correct option for comparison questions), taken from the training image, averaged over $\mathcal{H}$. $\mathcal{L}_{\text{fid}}$ is the KL divergence from the adapter-off to the adapter-on model at the first token of a prompt that names no region, applied to every fourth item. Organ boxes enter only the training loss, and no region is given at test.

Recipe and variants. Training uses half of SLAKE's images, with closed English questions, for 400 steps of 8 items, AdamW with learning rate 2×10^{-4}, no weight decay, gradient clipping at 1.0 and seed 42 (seed 7 on the second split), on 600 training items that mix organ questions with a box, which carry the attention term, and other closed questions, which carry the answer loss only, and evaluation uses up to 300 questions from the other half with no region given. A sweep of variants on three pilot models chose the recipe, and a second image split with 4 models confirmed it, as Appendix D.5 reports. The *reliance variant* raises the attention weight to 1 and adds a counterfactual hinge with margin 1 at weight 0.5, which requires the answer's log-probability to fall when the organ is mean-filled. The reliance variant in Table 1 trains for 200 steps. At 400 steps, the recipe's length, it adds 11.1 points of accuracy on the main split and 5.4 on the second, multiplies attention in the organ 4.6-fold, and still makes the answer depend on the organ in 7 of 7 models. It is the variant that makes the answer depend on the organ.

B.7 Report generation and the clinical labeller

Reports for 199 Chest ImaGenome studies are generated with 10 models under four conditions, (i) *frozen*, a free findings prompt, (ii) *MedGaze*, the same prompt with the adapter on and no region, (iii) the *unsteered region tour*, one sentence per anatomical region (right lung, left lung, cardiac silhouette, mediastinum, spine), each prompted by name and concatenated, with the heads left alone, and (iv) the *steered region tour*, the same prompts with the heads steered at $b = 4$ to that region's Chest ImaGenome box. The unsteered region tour separates steering from prompting, since naming a region is itself an intervention. The reference is the set of findings that the study's scene graph marks as present, and a report asserts a finding when a synonym appears in a sentence with no negation cue before it. Precision, recall and F1 run over positive findings only. The labeller is rule-based and reproducible, and every report comparison is a paired difference on identical studies. Lexical metrics are scored against the radiologist's report.

B.8 Evaluation protocols

Montage judge. An unsteered `Qwen3-VL-8B` reads the plain montage and the generated description and names the described panel, forced to choose one of six and never told the target. An output identical to the unsteered baseline counts as a miss.

Degeneracy filter. Steering can break the text into loops that leak content words, so the filter flags an output by length, alphabetic fraction, repetition and type-token ratio, and adds the share of repeated 4-grams, which catches clause loops. A degenerate output counts as a miss and is reported separately.

Visual reliance. The *visual reliance index* is the drop in the reference phrase's log-probability when a counterfactual replaces the image, with four counterfactuals, (i) a blank image, (ii) a different patient's image of the same modality and view, (iii) the same image with the annotated box blacked out, and (iv) the same image with its patches shuffled. The effect is Cohen's d over findings. The different-patient counterfactual is the headline, because the language prior confounds the blank image, as Appendix D.3 shows.

Organ VQA. Closed SLAKE questions that name an organ fall into two families, organ choice ("which organ is . . . ") and yes/no. Accuracy is reported for both, and the skewed yes/no family also reports the Matthews correlation coefficient (MCC), which a constant answer cannot inflate.

C Extended Experimental Details

Every experiment uses the models of Table 2 and the six corpora below. The evaluation runs 16 protocols in five task families over six corpora, 693 model runs that score 308,822 items, and every model enters discovery and ACG, as Figure 7 shows protocol by protocol. Its hub holds the 25 models, 2.1B to 73.4B parameters, with an arc joining each base to its medical tunes, and the inner ring names each task family and its metric. Each outer cell is one protocol with the number of models that ran it and the items they scored, and its lettered chips mark its corpora, coloured by modality. The cards give each corpus's protocols, runs and items, with one question and answer verbatim from each open corpus and one radiograph with its dataset boxes and label, and no report text, from each credentialed chest X-ray corpus. Unless stated, every cross-model statistic treats one model as one observation, and every number is regenerated from the run files that Appendix E describes.

Data. VQA-RAD (Lau et al., 2018), PathVQA (He et al., 2020) and SLAKE (Liu et al., 2021) are open, whereas MS-CXR (Boecking et al., 2022) on MIMIC-CXR (Johnson et al., 2019), Chest ImaGenome (Wu et al., 2021) and VinDr-CXR (Nguyen et al., 2022) are credentialed PhysioNet data. The credentialed corpora are processed on an approved cluster under their data use agreements, and no reference report text is reproduced. The montage corpora draw panels from pools of 1,793 VQA-RAD and 19,654 PathVQA images, and SLAKE carries single-expert organ boxes, MS-CXR a radiologist box per report phrase, Chest ImaGenome automatic scene graphs with a gold subset, and VinDr-CXR finding boxes agreed by at least three of 17 radiologists.

Discovery run. The run is identical for every model and tuned for none: 200 montages of six panels, 448 pixels high with 8-pixel gaps, six queries per montage for 1,200 forward passes, the attention row of the last prompt token, raw region mass, the top 100 heads and seed 42.

Metrics. Discovery reports fixation concentration, its ratio to random heads, top-head retention, cross-corpus Spearman ρ and anatomy enrichment; steering and grounding report steering accuracy, Cohen's d, win rate, image-swap d and the grounding gap; ACG reports IoU with the organ box, hit rate and the attention share in the mask. Organ VQA reports accuracy, MCC, balanced accuracy, rare-class F1, attention in the organ and heads peaking in it, and report generation reports clinical F1, precision and recall alongside BLEU-4, ROUGE-L, METEOR, CIDEr and BERTScore.

Models. Table 2 lists all 25 models, from 2B to 73B parameters, in 3 families, Qwen3-VL (Bai et al., 2025b), Qwen2.5-VL (Bai et al., 2025a) and Gemma3 (Team et al., 2025). The medical models include Lingshu (Xu et al., 2026), HuatuoGPT-Vision (Chen et al., 2024) and MedGemma (Sellergren et al., 2025), and the reasoning tunes MedVLThinker (Huang et al., 2025), QoQ-Med and MediX-R1 (Mullappilly et al., 2026). 14 medical tunes have their own base in the roster, so head indices correspond one to one. Three generalist sizes lie outside the 4B to 34B band and are marked as legacy.

Statistics. Paired comparisons across models use the Wilcoxon signed-rank test, and medical against generalist models use the Mann-Whitney test. Per-finding likelihood effects are Cohen's d. Per-model comparisons of paired answers and estimates pooled over questions use exact McNemar

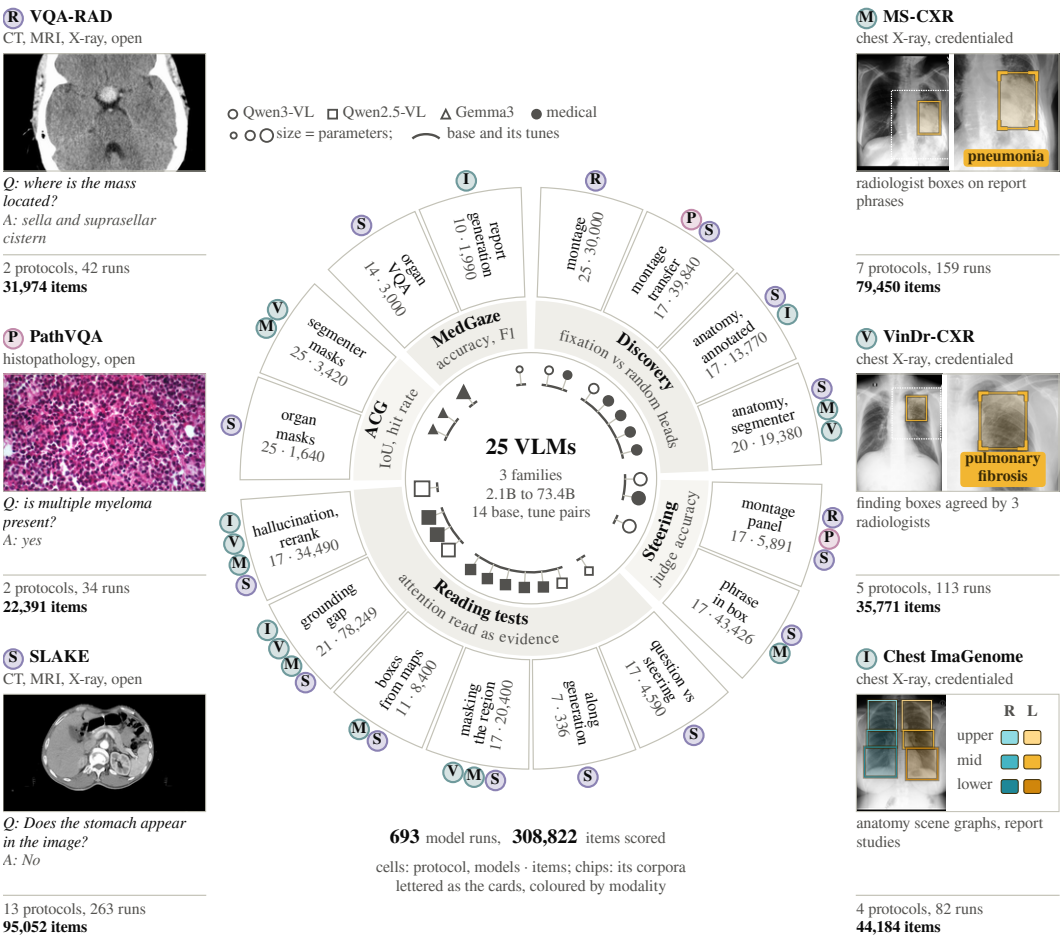


Figure 7: **Every one of the 25 models enters discovery and ACG, and the study scores 308,822 items in 693 model runs over 16 protocols and six corpora.** Hub: models (marker shape: family; filled: medical; size: parameters); rings: task families and protocols; cards: corpora.

tests on paired flips, and pooled estimates are always labelled "pooled over questions". Confidence intervals are bootstrapped over models, and Holm's correction runs within each family of tests, the metrics of one table or claim, such as the bottom rows of Table 9. The seed floor is the change between two seeds of the same recipe on 4 models. A non-significant result is reported with its p-value and never claimed as an effect.

D Extended Results and Analysis

The extended results follow Section 5 one result at a time, in the order of the main text.

D.1 Fixation heads exist and are inherited

Every one of the 25 models has fixation heads, and the per-model and per-tune evidence behind Section 5.1 follows. Each model's heads, their overlap with the base, their enrichment in radiologist boxes and their fixation matrices are collected in Figure 8, whose top row shows the montage input on `Lingshu-7B`, six unrelated VQA-RAD images queried by position.

Every model has fixation heads. The top-100 heads of `Lingshu-7B` put 0.36 of their image attention on the queried panel, whereas 100 random heads outside the top 100 stay near chance, at 0.18 against 0.167, as the top row of Figure 8 shows. The top head scores at least twice chance in 21 of 25 models, up to $5.0\times$ in `Qwen3-VL-2B`, as Figure 8a shows. Where the top head clears chance, it sits at 0.60 to

Table 2: **The fixation heads concentrate on the named panel above 100 random heads in all 25 models, and every supervised Qwen tune paired with its base keeps the base's top head.** Models by family († medical; legacy sizes last): top-100-head concentration (blue: 0.19 to 0.62), its change from the base, the same score for 100 random heads, and whether the base's top head is kept (✓; 4 of 4 supervised tunes).

		architecture			concentration			
Model	Base	params	$L \times H$	top head	fixation ↑	vs base	random	depth
Generalist bases								
Qwen2.5-VL-7B	n/a	8.3B	28×28	L19 H22	0.337		0.187	76%
Qwen2.5-VL-32B	n/a	33.5B	64×40	L52 H10	0.317		0.172	88%
Qwen3-VL-4B	n/a	4.4B	36×32	L23 H10	0.540		0.182	68%
Qwen3-VL-8B	n/a	8.8B	36×32	L23 H10	0.607		0.190	65%
Qwen3-VL-32B	n/a	33.4B	64×64	L48 H37	0.365		0.181	79%
Qwen3-VL-30B-A3B	n/a	31.1B	48×32	L40 H14	0.511		0.211	79%
Medical, supervised fine-tuning								
MedGemma-4B†	n/a	4.3B	34×8	L24 H4	0.202		0.169	68%
MedGemma-27B†	n/a	28.8B	62×32	L53 H0	0.230		0.179	61%
Lingshu-7B†	Qwen2.5-VL-7B	8.3B	28×28	L19 H22 ✓	0.362	+7.6%	0.183	77%
HuatuoGPT-Vision-7B†	Qwen2.5-VL-7B	8.3B	28×28	L19 H22 ✓	0.317	−5.8%	0.178	76%
Lingshu-32B†	Qwen2.5-VL-32B	33.5B	64×40	L52 H10 ✓	0.385	+21.5%	0.174	79%
Qwen3-VL-8B-MedSFT†	Qwen3-VL-8B	8.8B	36×32	L23 H10 ✓	**0.621**	+2.2%	0.186	65%
Medical, reasoning (RL / CoT)								
MedVLThinker-7B†	Qwen2.5-VL-7B	8.3B	28×28	L21 H5 ×	0.369	+9.4%	0.181	77%
QoQ-Med-VL-7B†	Qwen2.5-VL-7B	8.3B	28×28	L21 H5 ×	0.312	−7.4%	0.183	76%
MedVLThinker-32B†	Qwen2.5-VL-32B	33.5B	64×40	L52 H10 ✓	0.314	−1.0%	0.171	86%
QoQ-Med3-VL-8B†	Qwen3-VL-8B	8.8B	36×32	L24 H23 ×	0.604	−0.5%	0.193	65%
MediX-R1-8B†	Qwen3-VL-8B	8.8B	36×32	L24 H23 ×	0.611	+0.6%	0.206	65%
MediX-R1-30B†	Qwen3-VL-30B-A3B	31.1B	48×32	L0 H15 ×	0.200	−60.9%	0.178	55%
Medical, grounding-trained								
MedGemma-1.5-4B†	MedGemma-4B	4.3B	34×8	L17 H0 ×	0.190	−6.4%	0.169	60%
UniReason-Med†	Qwen2.5-VL-7B	8.3B	28×28	L21 H5 ×	0.348	+3.3%	0.185	76%
MedMO-4B†	Qwen3-VL-4B	4.4B	36×32	L24 H13 ×	0.308	−43.0%	0.180	68%
MedMO-8B†	Qwen3-VL-8B	8.8B	36×32	L21 H18 ×	0.189	−68.8%	0.178	35%
Generalist bases, legacy sizes (outside the 4B to 34B band)								
Qwen2.5-VL-3B	n/a	3.8B	36×16	L28 H10	0.368		0.179	80%
Qwen2.5-VL-72B	n/a	73.4B	80×64	L59 H21	0.493		0.199	77%
Qwen3-VL-2B	n/a	2.1B	28×16	L19 H5	0.501		0.190	67%

0.85 of network depth in Qwen3-VL (10 of 11 models) and at 0.70 to 0.83 in Qwen2.5-VL. Family orders the models more than size does, with a median of 2.20× chance in the 22 Qwen-VL models against 1.21× in the 3 Gemma3 models.

Medical tunes inherit them. Over 14 pairs, a tune's median share of its base's top-k heads is 0.80, 0.74, 0.80 and 0.81 at $k = 10$, 25, 50 and 100, against at most 0.09 by chance, as Figure 8b shows. All 4 supervised Qwen tunes keep the base's top head, against 1 of 10 other tunes, which are mostly reinforcement-learning and reasoning tunes. Each of the 7 Qwen2.5-VL tunes has its base's first or second head as its top head.

Montage heads ground on real boxes. The montage-discovered top-100 heads, never selected on MS-CXR, put more attention in the radiologist's box than in a control box in 11 of 17 models, with `Lingshu-7B` at 1.60 against 0.96 in the median enrichment of Figure 8c.

The fixation matrix is inherited. Over 14 pairs, the fixation matrix of a base's top head and that of the same head in its tune correlate at $r = 0.91$ on average, against 0.42 for mismatched heads (1.2×10^{-4}), as Figure 8d shows. Across all 25 models the top head's diagonal concentration is 0.62, against 0.17 for random heads. The inherited head also follows real anatomy, since L19·H22, the top head of `Qwen2.5-VL-7B`, is the top head on segmented SLAKE CT and MS-CXR organs in the base and in `Lingshu-7B`, as Figure 8e shows.

Stable across corpora, cut-offs and protocols. The same heads rank highest on radiology, pathology and mixed-modality montages, with Spearman $\rho = 0.96$ and the same top head in 13 of 17 models. All 25 of 25 models stay above chance at every cut-off from 1% to 25% of heads. Naming the anatomical region instead of its position raises concentration on every corpus, and on segmented anatomy named queries beat positional ones in 18 of 19 models ($p = 7 \times 10^{-5}$). Segmenter concentration correlates with the MS-CXR grounding gap ($\rho = +0.54$, $p = 0.033$), whereas montage concentration does not ($\rho = +0.042$, $p = 0.87$).

Family orders the models more than scale. The specialisation index, concentration in the top 1% of heads divided by that in the top 25%, is 1.63 to 2.14 in the Qwen-VL models and 1.20 to 1.36 in Gemma3. Gemma3 encodes every image into a fixed grid of 256 tokens, yet `MedGemma-4B` stays near

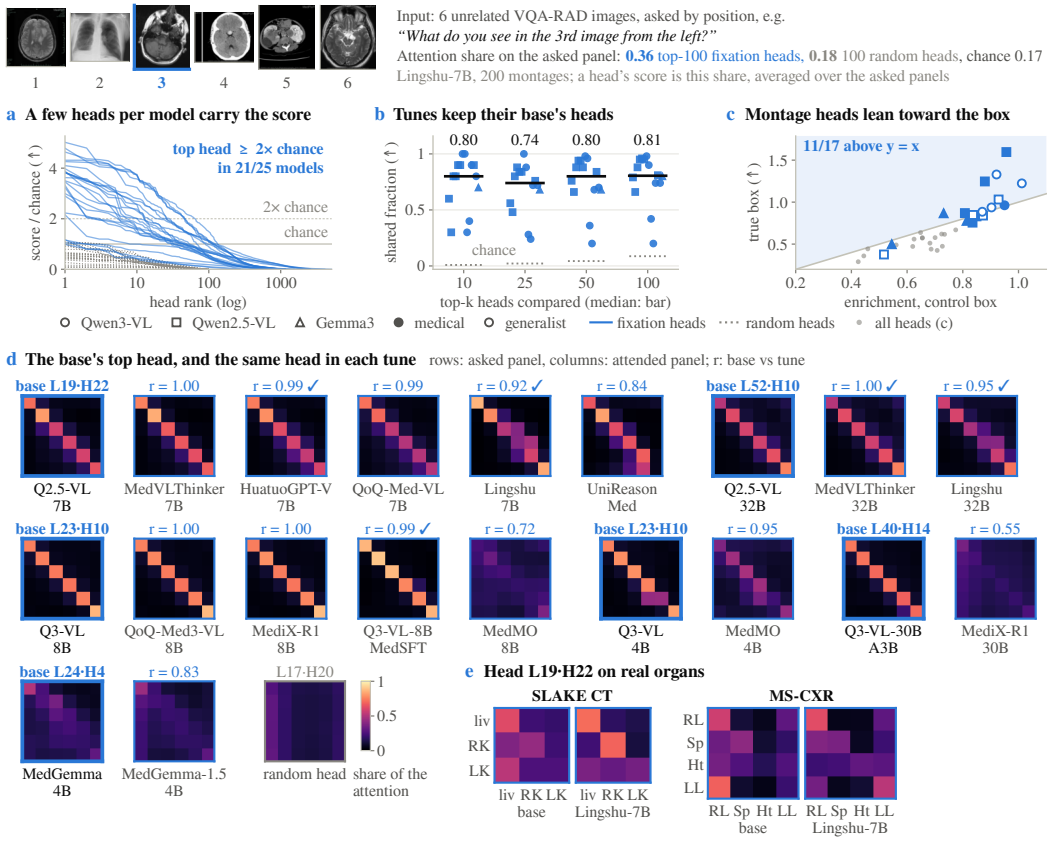


Figure 8: **Every model has fixation heads, and its medical tunes inherit them.** Top: a montage on `Lingshu-7B`; (a) score over chance by head rank; (b) top-k overlap with the base; (c) MS-CXR box enrichment; (d) fixation matrices; (e) L19·H22 on segmented anatomy.

chance with two and three montage panels, whereas `Qwen3-VL-8B` separates further from chance as panels are added, so the Gemma3 pattern is not an artefact of the token budget.

Gemma3 generalist bases. The roster's Gemma3 models are all medical, so their generalist bases, `Gemma-3-4B` and `Gemma-3-27B`, were run with the same discovery protocol and paired with their tunes, as Table 3 reports. The 4B base is as diffuse as its tunes (0.183 against a chance level of 0.167, tunes +11% and +4%), whereas the 27B base reaches 0.304 and `MedGemma-27B` loses 24% of it. On VQA-RAD the top-10 and top-100 overlaps are 0.50 and 0.79 for `MedGemma-4B`, 0.60 and 0.70 for `MedGemma-1.5-4B` and 0.30 and 0.42 for `MedGemma-27B`, against chance levels of 0.04 and 0.37 at 4B and 0.005 and 0.05 at 27B (hypergeometric $p \leq 1.1\times10^{-5}$ for every pair and k; over all 17 pairs, one-sided Wilcoxon $p_{\text{Holm}} \leq 5.8\times10^{-4}$). `MedGemma-1.5-4B` keeps the base's top head, which `MedGemma-4B` ranks third and `MedGemma-27B` 31st, and each Gemma3 base agrees with itself across corpora at 0.86 and 0.87 (top-100 overlap, VQA-RAD against SLAKE).

D.2 The heads are steerable

Writing a region into the fixation heads changes what the model says, and random heads steered the same way do much less. Each model's steering effect is one point in Figure 9, with random heads on the x-axis and fixation heads on the y-axis, so a point above the diagonal favours the fixation heads.

Steering redirects the description. Steered to a montage panel, the fixation heads make the judge pick that panel in 37.2% of cases against 21.7% for as many random heads ($p = 0.0017$), and they lead in 14 of 17 models. The test replicates on PathVQA montages, where fixation heads lead in 15/17 models ($p = 0.00094$), and on SLAKE montages (15/17, $p = 0.00029$), as Figure 9a and b

Table 3: **Every MedGemma tune keeps more of its Gemma3 base's fixation heads than chance on all three corpora, with a top-100 overlap of 1.8 to 8.9 times chance.** Gemma3 bases against MedGemma tunes, roster protocol. Top head: kept (✓) or not, with the rank of the base's top head in the tune; overlap shaded by its multiple of chance; bold: largest overlap per corpus.

		concentration				top-k overlap with base				chance
Tune	Base	base	tune	change	top head	k=10 ↑	25 ↑	50 ↑	100 ↑	k=100
Gemma3 pairs, VQA-RAD montages										
MedGemma-4B	Gemma-3-4B	0.183	0.202	+11%	× 3	0.50	0.44	0.64	0.79	0.37
MedGemma-1.5-4B	Gemma-3-4B	0.183	0.190	+4%	✓ 1	0.60	0.56	0.62	0.70	0.37
MedGemma-27B	Gemma-3-27B	0.304	0.230	−24%	× 31	0.30	0.32	0.36	0.42	0.05
MedGemma-1.5-4B	MedGemma-4B	0.202	0.190	−6%	× 3	**0.70**	**0.68**	**0.68**	**0.80**	0.37
Gemma3 pairs, SLAKE montages										
MedGemma-4B	Gemma-3-4B	0.186	0.201	+8%	× 2	0.50	0.60	0.68	0.74	0.37
MedGemma-1.5-4B	Gemma-3-4B	0.186	0.193	+4%	✓ 1	0.60	0.56	0.62	0.71	0.37
MedGemma-27B	Gemma-3-27B	0.321	0.250	−22%	× 17	0.20	0.40	0.34	0.43	0.05
MedGemma-1.5-4B	MedGemma-4B	0.201	0.193	−4%	× 2	**0.70**	**0.68**	**0.70**	**0.83**	0.37
Gemma3 pairs, PathVQA montages										
MedGemma-4B	Gemma-3-4B	0.190	0.203	+7%	× 3	0.20	0.52	0.62	0.72	0.37
MedGemma-1.5-4B	Gemma-3-4B	0.190	0.193	+2%	✓ 1	**0.60**	0.52	0.58	0.67	0.37
MedGemma-27B	Gemma-3-27B	0.378	0.256	−32%	× 17	0.10	0.32	0.38	0.45	0.05
MedGemma-1.5-4B	MedGemma-4B	0.203	0.193	−5%	× 2	**0.60**	**0.64**	**0.70**	**0.80**	0.37
Qwen pairs of the roster, VQA-RAD montages (13 pairs)										
median	n/a			−0.5%	5/13	0.80	0.76	0.84	0.81	≤0.13
min, max	n/a			−69, +22%		0.30, 1.00	0.24, 1.00	0.20, 0.98	0.20, 0.98	

show. Steering every head instead collapses accuracy to 1.3%, with 88% to 100% degenerate output per model.

Steering works on real anatomy. Steered to the radiologist's box on MS-CXR, the fixation heads raise the phrase's likelihood with $d = +0.36$, against $+0.08$ for random heads, and lead in 14 of 17 models ($p = 0.0079$). Steered to SLAKE boxes, organs and lesions, they reach $d = +0.53$ against $+0.13$ and lead in 15 of 17 models ($p = 0.00021$), as Figure 9c shows. A different patient's image lowers the phrase's likelihood in 17 of 17 models on MS-CXR, and on SLAKE it lowers it in all 17 models, as Figure 9d shows.

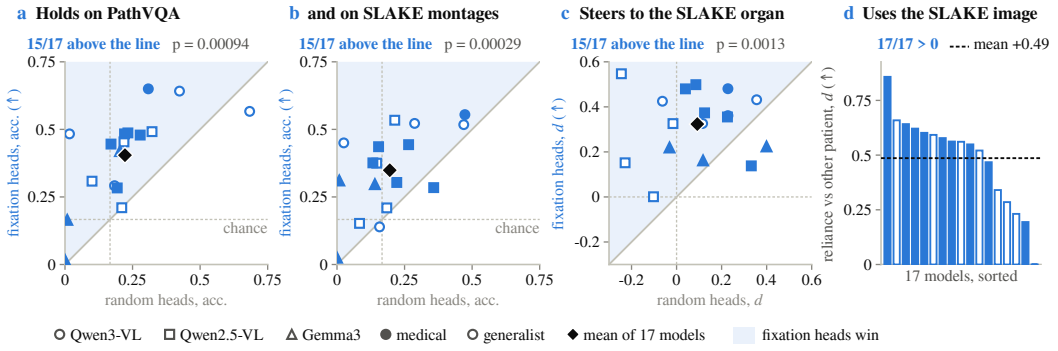


Figure 9: **Fixation heads beat random heads on the corpora the main text leaves out.** Random heads (x) against fixation heads (y) for montage accuracy on (a) PathVQA and (b) SLAKE and (c) Cohen's d on SLAKE boxes; (d) reliance on the SLAKE image per model.

Scoring every phrase token. The box-steering values above run the last prompt token as a steered step, as Appendix B.3 describes, so every phrase token is scored. Scoring the phrase from its second token instead, as an ordinary prefill does, gives $d = 0.34$ on MS-CXR against 0.36 with the steered step, which is significant in 14 of 17 models. On SLAKE, where labels are one-word organs, the steered step raises d from 0.32 to 0.53, significant in 16 of 17 models, and fixation heads beat random heads in 15 of 17 models ($p = 0.00021$).

Choosing the strength. Table 4 trades localisation against fluency over the bias b. At b=4 the heads already put 0.97 of their attention inside the steered region and no output is degenerate; from b=8 localisation saturates while degenerate output climbs to 35%. The hit rate is the share of steered sentences naming the region, against a chance of 0.11. Generation uses $b = 4$ except in the montage test, which keeps the reference recipe's hard gate $b = 10^4$, as does scoring a fixed phrase.

Table 4: **At b=4 the heads already sit in the region with no degenerate text.** Steering strength, 3 models on SLAKE (25 images × 9 regions); blue: larger is darker.

bias b	1	2	**4**	8	16	32	64
degenerate output ↓	0%	0%	0%	6%	35%	34%	35%
hit rate ↑	0.27	0.32	0.32	0.37	0.36	0.37	0.37
specificity ↑	0.47	0.66	0.78	0.82	0.82	0.82	0.82
attn. in region ↑	0.42	0.79	0.97	0.99	0.99	0.99	0.99

Robust to head count, bias, mode and control box. For the two models with a clear default effect, Lingshu-7B and Qwen3-VL-2B, all 14 of 14 alternative settings of head count, bias and mode keep $d > 0$ with $p < 0.05$ ($d = 0.24$ to 0.57), and the default sits at or near the best. Against a box shifted into the same neighbourhood, 3 of the 4 models with a significant grounding gap against a random box keep it. `Lingshu-7B` has $1.86\times$ enrichment in the box against $0.95\times$ for the shifted box and $0.87\times$ for the random box ($p = 8 \times 10^{-28}$, 250 findings).

Per-model variation. `Qwen2.5-VL-7B`, with no effect on MS-CXR ($d = -0.05$), steers strongly on SLAKE ($d = +0.53$). How strongly a model concentrates in discovery does not predict how well it steers ($\rho = +0.400$, $p = 0.11$).

D.3 The heads are not readable

A worked example first shows why the heads' attention cannot be read as the region the model used, and the paragraphs after it follow the five tests of Section 5.3 in the same order.

The problem in one example. Asked about the liver of an abdominal CT scan, `Lingshu-7B` answers that the liver looks normal, while its fixation heads put only 3% of their image attention on the liver and more on the small bowel, as Figure 10a shows. The answer therefore names an organ that the heads do not attend to. Asking the same scan about three different organs gives three nearly identical attention maps (cosine similarity 0.92 to 0.96), so the map does not follow the question, as Figure 10b shows. The pattern holds across all 17 models, where maps for different questions on one image have a correlation of 0.916, against 0.417 for one question on different images.

A geometric view. Figure 10c places each attention map in a triangle whose corners are three organs of the scan, at the point given by the attention share on each organ renormalised over the three. Changing the question moves the frozen model's point by only 0.12 in total variation, so the map is not readable. Steering the fixation heads to one organ sends the point to that organ's corner, a move of 0.55, so the heads are steerable. Over 210 questions on 35 scans, each seen by two models, the asked organ receives a raw share of 0.03 of the attention in the frozen model and 0.54 when the region is written. MedGaze, with no region given, raises this share from 0.03 to 0.12 and moves attention onto the anatomy as a whole.

Attention barely follows the question. Asked about three organs of one CT scan, the fixation-head map barely changes, as Figure 10b shows. Over 17 models, maps for different questions on one image have a correlation of 0.916 against 0.417 across images, and the within-image correlation is higher in 17 of 17 models ($p = 2 \times 10^{-5}$). On 35 SLAKE abdominal scans with liver and kidney masks, asking about an organ instead of another raises the heads' share of image attention on it by only 0.67 ± 0.70 points over 17 models. Steering the same heads to that organ raises the share by 53.8 ± 3.2 points, $80\times$ the question effect and larger in all 17 models (Holm $p = 1.5 \times 10^{-4}$).

Attention leaves the organ as the model answers. Asked to reason step by step, every model's fixation heads put more than the organ's token share on the organ while reasoning and fall back when writing the answer word. The organ share peaks at 0.43 of the reasoning span on average and falls on the answer in 295 of 336 questions (pooled over questions, $p = 5.9 \times 10^{-44}$), and each of the 7 models shows this fall in 33 to 47 of its 48 questions, as Figure 11 shows. There the share is taken inside the organ box, chance is the box's share of image tokens, and the bottom row of panel a is the mean over models. Panel b marks each model's median with a black tick, counts these questions at its right, pools all 336 questions in its bottom row with a box over the interquartile range, and places one `Qwen3-VL-8B` question at -31 points at its left edge. Averaged over 11 models, steering holds the share at 0.93 throughout generation, random heads leave it at 0.09, and MedGaze lifts it to 0.25 with no region given.

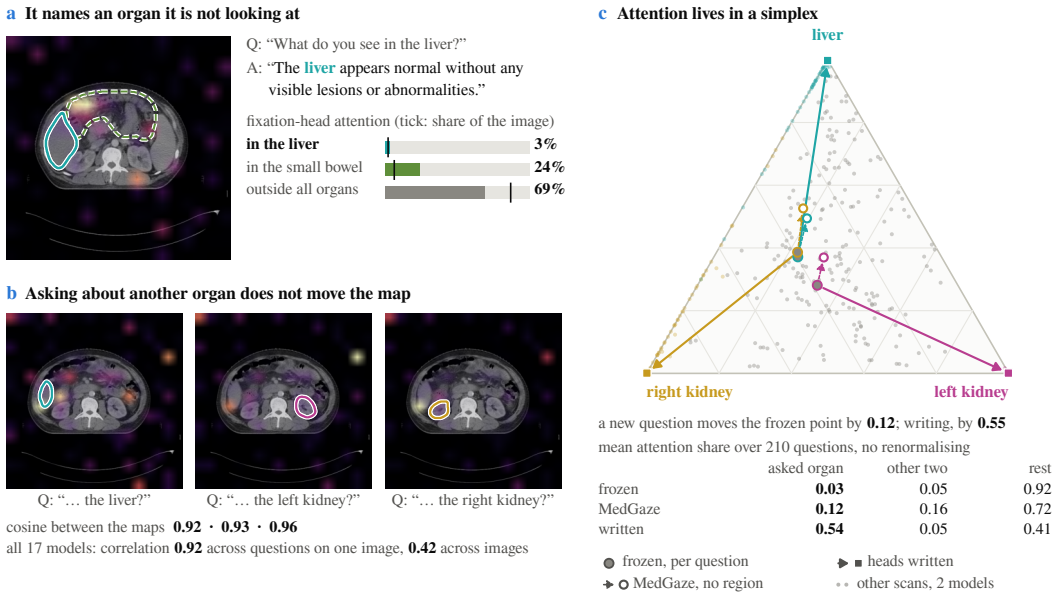


Figure 10: **The frozen attention map barely follows the question, whereas writing the region moves attention onto the asked organ.** `Lingshu-7B` on one SLAKE CT scan: (a) answer and map; (b) three organ questions; (c) the triangle of three organs.

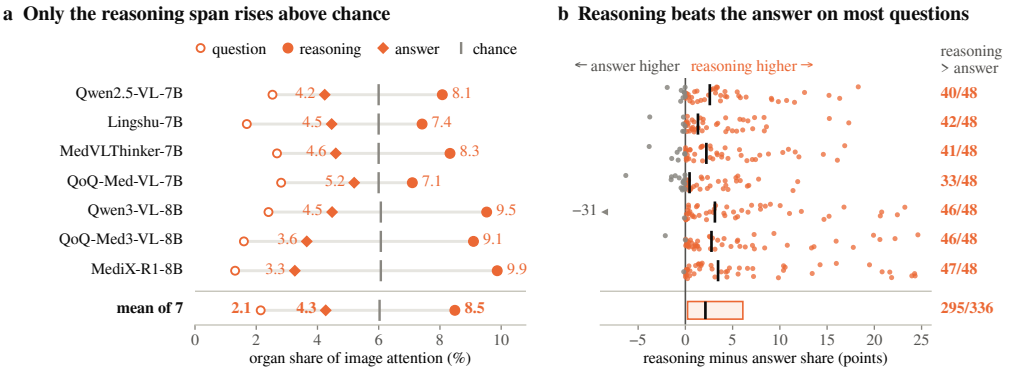


Figure 11: **In every model, attention on the organ rises above chance while the model reasons and falls below it on the answer.** (a) Organ share at the question, while reasoning and at the answer (grey tick: chance); (b) reasoning minus answer per question (orange: reasoning higher).

Masking the region leaves the phrase unchanged. Blacking out the radiologist's box leaves the phrase's likelihood unchanged, with mean $d \approx 0$ over 17 models against $+0.26$ for a different patient's image, as Table 5 lists per model. Shuffling the patches has no effect on average either ($d \approx 0$), whereas the blank image is confounded by the language prior and reverses sign in 9 of 17 models. On SLAKE and VinDr-CXR, blacking out the box leaves the likelihood unchanged as well (mean d of $+0.11$ and $+0.14$, Holm $p = 1.00$), whereas another patient's image lowers it more than the masked box in 16 of 17 models on SLAKE ($p = 8.4 \times 10^{-4}$). The contrast is sharpest in the 8 models whose phrase depends most on the image (other-patient $d \geq 0.2$ on MS-CXR), where another patient's image lowers the likelihood more than the masked box in all 8 ($d = 0.45$ against 0.09, $p = 0.008$). In the 8 models with a measurable baseline ($d \geq 0.2$), ablating the 100 fixation heads keeps 82% of this reliance on the image and ablating 100 random heads keeps 82% ($p = 0.95$), whereas ablating every head removes it.

Table 5: **Another patient's image lowers the phrase in every model, whereas masking the radiologist's box barely moves it (mean d +0.26 against 0.00).** Visual reliance on MS-CXR under four counterfactuals, and the share kept after ablating 100 fixation or random heads (models with other-patient $d \geq 0.2$; n/a below it). Green/orange: sign.

	visual reliance, Cohen's d of the phrase log-probability drop				reliance kept after ablation	
Model	blank image	other patient	box masked	patches shuffled	fixation heads	random heads
Lingshu-32B†	−0.08	+0.63	+0.42	+1.48	94%	100%
Lingshu-7B†	+1.18	+0.60	+0.36	+1.54	44%	96%
MedGemma-1.5-4B†	+0.51	+0.59	+0.14	−0.28	99%	66%
MedGemma-4B†	+0.94	+0.54	+0.08	+0.26	54%	25%
MedGemma-27B†	+0.42	+0.44	−0.05	−0.03	82%	85%
Qwen3-VL-32B	+0.62	+0.30	−0.18	−0.42	97%	91%
Qwen3-VL-2B	+1.29	+0.25	−0.16	+0.66	79%	81%
Qwen3-VL-4B	−0.69	+0.24	+0.12	−0.41	104%	110%
Qwen3-VL-8B	−1.07	+0.20	−0.24	−0.74	n/a	n/a
QoQ-Med3-VL-8B†	−1.02	+0.20	−0.28	−0.83	n/a	n/a
Qwen2.5-VL-72B	+0.48	+0.11	−0.14	+0.42	n/a	n/a
Qwen2.5-VL-32B	−1.69	+0.10	−0.14	−0.64	n/a	n/a
Qwen2.5-VL-7B	−0.62	+0.09	+0.16	−0.12	n/a	n/a
MedVLThinker-7B†	−0.64	+0.08	−0.02	−0.19	n/a	n/a
HuatuoGPT-Vision-7B†	−0.28	+0.07	−0.32	−0.70	n/a	n/a
Qwen2.5-VL-3B	+1.01	+0.01	+0.61	+0.82	n/a	n/a
QoQ-Med-VL-7B†	−0.58	+0.01	−0.40	−0.91	n/a	n/a
mean over 17 models	**−0.01**	**+0.26**	**0.00**	**−0.00**	**82%**	**82%**

Attention proposes no better regions than chance. Boxes read from the fixation-head map of a held-out proposer, `Qwen3-VL-32B`, on 400 MS-CXR findings reach a mean best IoU of 0.167 at the best threshold, against 0.188 for size-matched random boxes and 0.193 for a 3×2 grid ($p = 0.014$), and they lose at all 3 thresholds and with all heads. The map piles into the image corners, yet border attention is only $1.10\times$ uniform over 17 models (0.84 to 1.56), so the corners do not reflect a border prior. Standard attribution maps fail the same test, with the findings, prompt, box rule (percentile 60, up to six boxes) and scoring fixed, and Table 6 compares the mean over all heads, attention rollout (Abnar & Zuidema, 2020), gradient×attention relevance (Chefer et al., 2021) and a Grad-CAM-style input-gradient saliency (Selvaraju et al., 2017) with random boxes of their own sizes over 10 models. On MS-CXR no map beats its random boxes after Holm correction, the closest being gradient×attention at 0.174 ± 0.014 against 0.162 (+7% relative, 8 of 10 models, $p = 0.15$), and rollout falls below random (0.080 ± 0.022, $p = 0.014$). On 400 SLAKE organ boxes, 4 of 5 maps lose to their random boxes after correction and none wins in more than 1 of 10 models. A fixed 3×3 grid beats every map on both corpora, with IoU 0.239 on MS-CXR and 0.167 on SLAKE.

Table 6: **No attribution map proposes better boxes than random boxes of its own sizes, and a fixed 3×3 grid beats every map on both corpora.** Best IoU with the reference box, mean ± sd over models, against size-matched random boxes; wins: models above random; paired Wilcoxon, Holm within corpus; green beats random, orange loses; bold: best per corpus.

	MS-CXR (10 models, 400 findings)					SLAKE organs (10 models, 400 findings)				
Box proposer	IoU↑	random	hit@0.1↑	wins	p_{Holm}	IoU↑	random	hit@0.1↑	wins	p_{Holm}
fixation heads	0.154 ± 0.036	0.173	57%	1/10	0.068	0.078 ± 0.009	0.096	26%	1/10	0.055
all heads, mean	0.133 ± 0.023	0.154	50%	1/10	0.078	0.082 ± 0.017	0.105	27%	0/10	0.014
attention rollout	0.080 ± 0.022	0.099	26%	0/10	0.014	0.055 ± 0.009	0.068	20%	0/10	0.014
gradient×attention	0.174 ± 0.014	0.162	61%	8/10	0.15	0.081 ± 0.013	0.098	28%	1/10	0.016
input gradient	0.154 ± 0.011	0.160	58%	4/10	0.28	0.091 ± 0.014	0.109	30%	1/10	0.016
random boxes, fixation-sized	0.173 ± 0.033		60%			0.096 ± 0.019		33%		
grid, 3×2	0.193	0.173	81%	8/10	0.15	0.102	0.096	37%	6/10	0.28
grid, 3×3	**0.239**	0.173	89%	10/10	0.014	**0.167**	0.096	54%	10/10	0.014

Attention predicts neither hallucination nor answer quality. As a hallucination detector, fixation-head focality reaches AUROC 0.528 over 68 runs, against 0.526 for random heads and 0.531 for log-probability. Reranking eight sampled answers by grounding beats log-probability in only 15 of 51 runs and falls below random selection on MS-CXR, with every number taken under a neutral prompt that does not state the finding.

Attention sink. Every model parks 0.19 to 0.60 of the fixation heads' attention on a sink token over 17 models, measured at its true position. Penalising the sink lifts image attention $1.5\times$, yet under a neutral prompt the phrase's gain is not significant ($+0.07$, $p = 0.43$) and box enrichment does not change ($+0.00$).

D.4 WRITING THE REGION WITH ANATOMY-CONDITIONED GENERATION

Anatomy-conditioned generation attaches one region to every sentence at inference, and its region sources, hit rates and provenance test are detailed here. The four region sources on one SLAKE study of `Lingshu-7B`, their per-model hit rates on chest radiographs and what each steered sentence says are collected in Figure 12, and Table 7 gives each source's hit rate, IoU and ceiling, so a source that localises well has an IoU close to that of the annotated box.

The heads stay inside the region, and only steering puts them there. Steered to an organ mask at $b = 4$, the fixation heads place 0.90 to 0.99 of their attention inside it, with a median of 0.98 over 151 steered generations in 11 models. Cheaper ways of marking the same region, compared on 120 SLAKE images (169 organ masks per model) and 60 VinDr-CXR images (830 anatomy masks per model) for 5 models, leave the heads near their unsteered share of 7.7% on SLAKE and 7.0% on VinDr-CXR. Naming the organ in the prompt gives 7.5% and 8.2%, a 2-pixel red outline of the mask 15.1% and 13.3%, and dimming the image outside the mask to 50% gives 15.1% and 12.3%, whereas ACG gives $94.5 \pm 1.3\%$ and $95.5 \pm 3.0\%$. Prompt and image cues change what the model is given, whereas steering changes where the heads look while the sentence is written.

Organ masks localise as well as the annotation. On SLAKE, organ masks select a region with IoU 0.830 against the annotated organ, close to the oracle box (0.919) and far above the 3×3 grid (0.075). Masks beat the grid in all 25 models ($p = 6 \times 10^{-8}$), whereas class-agnostic Segment Anything masks rarely follow organ boundaries, which caps their ceiling IoU.

Chest regions need no annotation. On chest radiographs, anatomy masks from a segmenter run on the test image select the finding's region at $1.8\times$ chance on MS-CXR and $2.0\times$ on VinDr-CXR, as Figure 12b shows. Checked before any model runs, the segmenter's regions contain the annotated box more often and with a higher best IoU than a 3×3 grid on both chest corpora. The grid reaches a higher raw hit rate with its fewer and larger cells, whereas its IoU stays low.

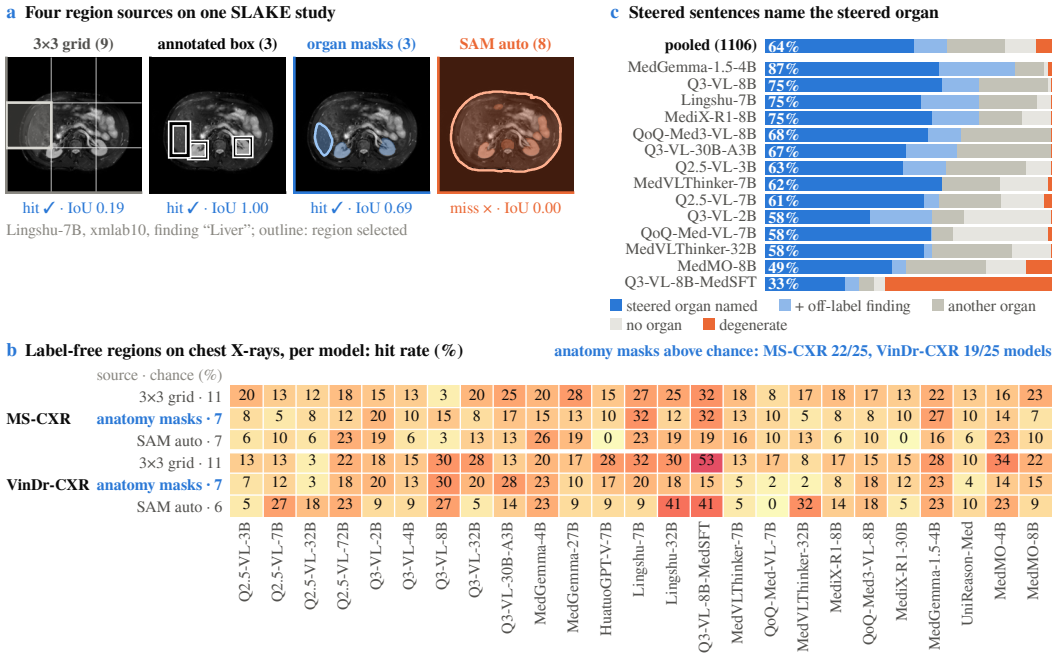


Figure 12: **Label-free regions steer the heads, and most steered sentences name the steered organ.** (a) The four region sources on one SLAKE study; (b) hit rate per model on MS-CXR and VinDr-CXR; (c) what each steered sentence says, per model and pooled.

Most steered sentences name the steered organ. Of 1106 sentences generated in 14 models with the heads steered to an organ mask, 64% name the steered organ, as the rule-based classification in Figure 12c shows. Figure 14 in Appendix F shows ACG on four more SLAKE studies.

Table 7: **Segmentation masks give ACG its tightest region without annotation on SLAKE and VinDr-CXR, and on SLAKE they come close to the annotated box (IoU 0.830 against 0.919).** Anatomy-conditioned generation by region source, mean over models: hit, its chance, lenient hit, IoU of the chosen region and the best IoU the source allows; bold: best IoU without annotation per corpus.

Corpus (models)	Region source	hit ↑	chance	÷ chance ↑	lenient hit ↑	IoU ↑	ceiling IoU
SLAKE (25)	3×3 grid	0.173	0.111	1.6×	0.446	0.075	0.242
	organ masks (ours)	0.919	0.880	1.0×	0.939	**0.830**	0.892
	SAM, automatic	0.210	0.090	2.3×	0.343	0.038	0.160
	annotated box (oracle)	0.915	0.880	1.0×	0.937	0.919	1.000
MS-CXR (25)	3×3 grid	0.181	0.111	1.6×	0.422	0.060	0.221
	anatomy masks (ours)	0.131	0.073	1.8×	0.470	**0.063**	0.277
	SAM, automatic	0.127	0.071	1.8×	0.213	0.009	0.037
	annotated box (oracle)	0.893	0.872	1.0×	0.898	0.893	1.000
VinDr-CXR (25)	3×3 grid	0.206	0.111	1.9×	0.426	0.044	0.148
	anatomy masks (ours)	0.143	0.072	2.0×	0.550	**0.127**	0.457
	SAM, automatic	0.162	0.065	2.5×	0.188	0.005	0.016
	annotated box (oracle)	0.845	0.811	1.0×	0.851	0.845	1.000

The sentence comes from the steered region. To test provenance directly, the steered region is blanked, filled with the image mean, while the heads stay steered to it, and blanking a random region of the same size serves as the control. On SLAKE, with 454 steered regions in each of 8 models, blanking the steered region lowers the share of these sentences that name the organ from 52% to 22%, against 50% for the random region, in all 8 models ($p = 0.008$). In the steered report tour over 199 Chest ImaGenome studies, blanking each stop's box while that stop is described lowers finding F1 from 30.6 to 16.7 in 8 of 9 models ($p = 0.008$), whereas a random box of the same size gives 29.4. Without steering, blanking the same box moves F1 only from 21.9 to 17.3, as the masking test of Appendix D.3 predicts, so the box shapes the report once it is written into the heads (a difference of 9.2 points, $p = 0.012$). Table 8 lists every model, except `MedMO-8B` in the tour, as it writes no finding the labeller scores.

Table 8: **Blanking the steered region removes what the steered sentence says, and blanking a random region of the same size does not.** Left: ACG sentences on SLAKE that name the steered organ (%); right: finding F1 of the report tour on Chest ImaGenome.

	ACG, names the organ			Report tour, finding F1				
		organ	random	heads steered			not steered	
model	as is	blanked	blanked	as is	stop blanked	random blanked	as is	stop blanked
`HuatuoGPT-Vision-7B`	56	15	49			–		
`Lingshu-7B`	56	33	56	27.8	6.0	20.7	18.0	6.6
`MedGemma-1.5-4B`	62	27	59	49.4	26.8	50.4	38.7	23.2
`MediX-R1-8B`	53	22	57	29.3	14.0	32.5	24.9	23.8
`MedVLThinker-32B`		–		36.9	38.2	36.6	33.4	34.7
`MedVLThinker-7B`	48	12	37	31.1	25.3	31.2	27.6	20.2
`Qwen2.5-VL-7B`	43	15	39	26.8	18.0	26.3	19.9	19.4
`Qwen3-VL-2B`	45	16	45	25.6	0.7	28.2	2.9	3.9
`Qwen3-VL-8B`	56	33	56	31.0	17.2	29.9	19.6	20.0
`Qwen3-VL-8B-MedSFT`		–		17.6	4.5	9.2	12.2	3.5
mean	52	22	50	30.6	16.7	29.4	21.9	17.3

D.5 Learning to look with MedGaze

The heads can be trained to look at the anatomy, first with the region given, through calibrated gains, and then with none, through MedGaze. Table 9 lists each model's frozen and MedGaze values, where green cells improve on the frozen checkpoint.

Calibrated gains beat the fixed bias. With the region given, one learned gain per head adds $+1.5$ points pooled over questions and 15 models ($p = 3 \times 10^{-5}$, $n = 5880$), against $+1.3$ for the fixed bias ($p = 2 \times 10^{-7}$) and -0.1 for random heads. The calibration beats the fixed bias in 9 of 15 models on organ choice and 11 of 15 on yes/no. The fit is selective, as it turns 13% to 31% of the heads below a gain of 2 and spreads the rest up to 11.1.

MedGaze moves the heads onto the organ in every model. With no region at test, attention inside the named organ at the answer step rises from 0.09 to 0.33, and the share of heads peaking there, those whose strongest image token lies in the organ, rises from 0.10 to 0.45, both in 14 of 14 models. Accuracy rises from 75.1 to 84.8 ($+9.7$ points, mean over models) in 12 of 14 models, and yes/no MCC rises from 0.43 to 0.60 in 13 of 14. On yes/no questions MedGaze answers "yes" in 83% of cases, closer to the 85% of questions whose answer is yes than the frozen 68%, and balanced

accuracy rises in 8 of 14 models, while visual reliance moves from .35 to .23 (Holm $p = 0.0785$). Retrained with heads ranked on SLAKE training images only, MedGaze keeps its gain in all 7 retrained models (11.4 against 11.8 points, attention in the organ 0.29 against 0.30), so the held-out images play no role in the result. Each of these four changes clears the seed floor, the change between two seeds of one recipe, and survives Holm's correction, as the bottom rows of Table 9 show.

Table 9: **MedGaze raises attention in the organ in all 14 models and accuracy and MCC in 12 and 13 of them, each change above the seed floor and significant after Holm correction.** MG: MedGaze, fr.: frozen checkpoint, held-out SLAKE, no region at test; green/orange: better/worse; bold: best model; bottom rows: bootstrap 95% CI, seed floor and Holm-corrected Wilcoxon test over models.

	all organ questions								yes/no questions									
Model	in organ attn. ↑		in organ peak ↑		acc. ↑		visual rel.		acc. ↑		bal. acc. ↑		F1 rare ↑		MCC ↑		answ. "yes"	
	fr.	MG	fr.	MG	fr.	MG	fr.	MG	fr.	MG	fr.	MG	fr.	MG	fr.	MG	fr.	MG
Generalist																		
Qwen2.5-VL-7B	.061	.227	.04	.33	.633	.857	.10	.10	.633	.907	.785	.795	.444	.667	.404	.613	49%	87%
Qwen3-VL-30B	.127	.434	.21	.57	.724	.881	.40	.23	.727	.933	.821	.829	.506	.750	.462	.717	59%	88%
Qwen2.5-VL-3B	.072	.270	.07	.37	.762	.862	.10	.15	.780	.893	.852	.825	.560	.667	.521	.607	65%	83%
Qwen3-VL-2B	.112	.330	.16	.50	.771	.857	.46	.20	.813	.920	.665	.859	.417	.739	.308	.693	83%	84%
Qwen3-VL-8B	.100	.286	.11	.39	.767	.810	.31	.13	.773	.873	.773	.794	.500	.612	.419	.542	69%	82%
Medical																		
MedGemma-1.5†	.039	.295	.01	.41	.614	.814	.92	.42	.647	.853	.529	.839	.381	.621	.281	.561	60%	76%
QoQ-Med-VL-7B†	.058	.231	.04	.33	.629	.795	.15	.11	.607	.827	.713	.710	.392	.480	.302	.382	50%	81%
MedVLThink.-7B†	.065	.241	.05	.33	.676	.838	.09	.10	.700	.893	.805	.768	.483	.619	.436	.558	57%	87%
MedMO-8B†	.041	.283	.03	.38	.741	.826	.04	.03	.748	.867	.517	.748	.185	.576	.035	.497	85%	84%
QoQ-Med3-VL-8B†	.101	.291	.11	.39	.752	.829	.54	.16	.740	.893	.716	.825	.435	.667	.329	.607	69%	83%
Lingshu-7B†	.098	.480	.11	.58	.833	.876	.33	.39	.867	.900	.809	.828	.615	.681	.546	.624	80%	83%
MedVLThink.-32B†	.031	.304	.01	.39	.810	.852	.24	.23	.780	.860	.871	.843	.571	.632	.545	.573	63%	77%
MediX-R1-8B†	.114	.385	.13	.52	.848	.848	.58	.22	.860	.907	.824	.795	.618	.667	.553	.613	78%	87%
Qwen3-VL-8B-SFT†	.266	**.608**	.31	**.75**	.957	**.929**	.64	**.70**	.953	**.953**	.935	**.916**	.851	**.844**	.826	**.817**	83%	85%
mean over 14	.092	**.333**	.10	**.45**	.751	**.848**	.35	**.23**	.759	**.891**	.758	**.812**	.497	**.659**	.426	**.600**	68%	83%
models improved	14/14		14/14		12/14		5/14		13/14		8/14		13/14		13/14			
95% CI of the change	[.21, .28]		[.32, .38]		[.06, .14]		[−.22, −.03]								[.11, .25]			
seed floor	.015		.016		.053		.084								.131			
paired test, p_{Holm}	**<0.001**		**<0.001**		**0.0037**		0.0785								**<0.001**			

The attention term is what moves the heads. Table 10 lists every adapter variant on the pilot trio, with accuracy pooled over questions, below training-free steering with the organ mask given and above MedGaze on all 14 models. Its columns give the accuracy change in points, the trainable parameters, the attention on the organ with the adapter on divided by off, the heads peaking in it and the models whose visual reliance rises (paired Wilcoxon $p < 0.05$). The answer loss alone gains $+10.1$ points on the fixation heads and $+11.2$ on all heads, but it leaves attention on the organ unchanged ($1.0\times$). Only variants with the attention term move the heads onto the organ, by $4.1\times$ for MedGaze. The reliance variant adds a counterfactual term and makes the answer depend on the organ in 4 of 4 models on the second split.

Confirmation on a second split. Because the variants were selected on the split they are reported on, four of them were retrained on a second image split for 4 models. Matched for training steps, MedGaze gains $+8.3$ points and the answer-only adapter $+8.6$ (mean over models), so the attention term costs no accuracy. The change in attention replicates, $3.7\times$ for MedGaze against $1.1\times$ for the answer-only adapter.

Targeting the asked organ. The recipe extends to question-specific attention with a contrastive term that raises the heads' attention on the organs a question names relative to the other organs in the training image. This term makes yes/no attention more specific to the asked organ, as the asked-minus-other share rises from 6.8 to 8.1 points in 14 of 14 models ($p = 1.2 \times 10^{-4}$), with organ VQA accuracy unchanged (-0.3 points, $p = 0.54$).

The fixation heads are where the update belongs. The MedGaze recipe was also trained unchanged (rank 8, α=16, $\mathcal{L}_{\text{ans}} + 0.25\,\mathcal{L}_{\text{att}} + 0.1\,\mathcal{L}_{\text{fid}}$, 400 steps, same split and seed) on 100 seeded random heads outside the top 100, in each of the 14 models of Table 9. Scored on the heads each adapter trains, attention in the organ starts higher on the fixation heads (0.092 frozen, against 0.032 for the random heads) and ends higher, 0.333 ± 0.108 against 0.262 ± 0.065, in 14 of 14 models (paired Wilcoxon, Holm p<0.001). The fixation heads are therefore where the attention term puts the most attention on the organ.

Table 10: **The MedGaze recipe gives the largest accuracy gain of every adapter variant, and only variants with the attention term move attention onto the organ (2.8 to 9.6 times frozen, against 1.0 to 1.1 without it).** Top: steering with the organ mask, 15 models; middle: variants on the pilot trio (pooled, $n = 810$) and on a second split; bottom: all 14 models; attention: on ÷ off.

	Δ accuracy (pts) ↑			params	attention	heads	reliance
Variant	all	organ	yes/no	(M)	on organ ↑	peaking ↑	rises
region given, model frozen (15 models, all held-out images; pooled, $n = 5880$)							
random heads, fixed bias b=4	−0.1	+0.0	−0.2	0	n/a (the mask sets attention)		
fixation heads, fixed bias b=4	+1.3	+1.7	+1.1	0	n/a (the mask sets attention)		
fixation heads, one learned gain per head	+1.5	+2.2	+1.3	10^{-4}	n/a (the mask sets attention)		
with the attention term							
attention 0.25, 400 steps: MedGaze (ours)	+13.3	+8.3	+14.8	0.85	4.1×	0.41	0/3
reliance variant (attention 1, counterfactual 0.5)	+6.3	+5.6	+6.5	0.85	4.7×	0.46	3/3
reliance variant without counterfactual	+8.1	+6.1	+8.7	0.85	4.9×	0.48	0/3
attention 0.25	+10.0	+7.2	+10.8	0.85	4.0×	0.42	0/3
attention 0.1	+10.0	+6.7	+11.0	0.85	3.3×	0.36	1/3
attention 0.25, counterfactual 0.1	+9.1	+6.7	+9.8	0.85	4.0×	0.42	0/3
attention 0.25, top 300 heads	+12.7	+6.7	+14.4	1.11	4.3×	0.37	0/3
attention 0.25, open questions added	+9.6	+4.4	+11.1	0.85	2.8×	0.30	1/3
answer loss only							
fixation heads	+10.1	+6.7	+11.1	0.85	1.0×	0.08	0/3
all heads (unrestricted)	+11.2	+3.9	+13.3	1.61	1.1×	0.03	0/3
fixation heads, rank 16	+10.7	+6.7	+11.9	1.70	1.0×	0.08	0/3
top 300 heads	+11.2	+7.2	+12.4	1.11	1.0×	0.05	1/3
ablations of the reliance variant							
reliance variant without attention term	+8.8	+6.7	+9.4	0.85	1.0×	0.08	2/3
reliance variant on random heads	+7.2	+6.7	+7.3	0.91	9.6×	0.41	1/3
confirmation on a second image split (4 pilots, $n = 1136$; pooled over questions, then mean over models)							
MedGaze (ours; attention 0.25, 400 steps)	+8.9	mean over models +8.3			3.7×		0/4
reliance variant	+3.5	mean over models +3.2			4.1×		4/4
answer loss only	+7.1	mean over models +7.0			1.1×		0/4
answer loss only, 400 steps	+9.0	mean over models +8.6			1.1×		2/4
MedGaze on every model with an adapter (main split, 14 models; mean over models)							
MedGaze (ours)	+9.7	+0.6	+13.2	1.00	3.6×	0.45	1/14

MedGaze beats training-free decoders. Per model, exact McNemar tests on the paired answers favour MedGaze in 12 of 15 comparisons with the three decoders (VCD 3, PAI 5, CRG 4 of 5). MedGaze is the most accurate method in all 5 models and has the highest yes/no MCC in all 5, against three training-free decoders and fixation steering on the same 210 held-out SLAKE organ questions per model, as block A of Table 1 shows (paired Wilcoxon $p = 0.0625$ against each baseline, the smallest value possible with 5 models). Following the original papers, VCD (Leng et al., 2024) contrasts a diffusion-noised image ($\alpha = 1$, $\beta = 0.1$, noise step 500), PAI (Liu et al., 2024) adds $0.5|A|$ to the image attention logits of every head in all layers but the first two with an image-free reference at $\gamma = 1.1$, and CRG (Wan et al., 2024) contrasts the image with the organ mask blacked out ($\alpha = 1$), and all methods decode greedily and are scored on the first word. Above the frozen model, the other methods improve 4 (steering), 3 (VCD), 1 (CRG) and 0 (PAI) of the 5 models, whereas MedGaze raises yes/no balanced accuracy from $74.0 \pm 11.9\%$ to $80.5 \pm 2.9\%$ with the smallest spread across models of any method. PAI fails mainly on `MedGemma-1.5-4B` (16.7%), which writes a report instead of an answer, and looking for the answer anywhere in the text gives 17.6%, so this is not a scoring artefact.

D.6 Writing the region into reports

Free-text reports are the last setting: steering the heads region by region adds findings to the report, and MedGaze is compared on the same studies with no region at all. Table 11 lists each model's clinical finding F1, where green cells improve on the same model under the same prompt without the intervention.

Steering adds findings under the same prompt. Steering the heads over the region tour raises clinical finding F1 by $+7.85$ points against the same tour with the heads off, in 9 of 10 models (Holm $p = 0.023$).

MedGaze adds findings with no region. MedGaze alone raises F1 by $+2.78$ points in 8 of 10 models (uncorrected $p = 0.049$). Lexical metrics such as BLEU and ROUGE do not move, so the heads change which findings a report mentions rather than its wording. MedGaze also raises the log-likelihood of the reference caption of a named region in 8 of 10 models ($p = 0.014$).

Figure 16 in Appendix F shows report cards for both writes.

Table 11: **Steering the fixation heads region by region raises clinical finding F1 by 7.9 points in 9 of 10 models, significant after Holm correction, and MedGaze raises it by 2.8 points in 8 of 10 with no region given.** Chest ImaGenome, F1 in points († medical); shading as in Table 9; bottom rows: bootstrap CI and Wilcoxon test over models.

	region tour, identical prompt			no region given		
Model	heads off	**steered (ours)** ↑	Δ ↑	frozen	**MedGaze (ours)** ↑	Δ ↑
Lingshu-7B†	18.0	27.8	+9.8	19.8	24.9	+5.0
MedGemma-1.5-4B†	38.7	**49.4**	+10.8	35.0	35.2	+0.3
MedMO-8B†	0.0	0.0	+0.0	2.6	1.0	−1.6
MedVLThinker-32B†	33.4	36.9	+3.5	26.7	30.1	+3.3
MedVLThinker-7B†	27.6	31.1	+3.5	29.5	34.0	+4.6
MediX-R1-8B†	24.9	29.3	+4.5	39.0	**40.1**	+1.1
Qwen2.5-VL-7B	19.9	26.8	+6.9	23.5	28.0	+4.5
Qwen3-VL-2B	2.9	25.6	**+22.7**	10.3	22.5	**+12.2**
Qwen3-VL-8B	19.6	31.0	+11.4	38.5	35.3	−3.2
Qwen3-VL-8B-MedSFT†	12.2	17.6	+5.4	19.3	20.9	+1.6
macro mean, F1	19.7	**27.6**	**+7.9**	24.4	**27.2**	**+2.8**
precision	23.6	29.4	+5.8	28.1	30.3	+2.2
recall	21.7	32.7	+11.0	28.0	31.9	+4.0
paired over models: 95% CI of the change (points), models improved, p_{Holm} over the six tests						
F1	[4.5, 12.0]	9/10	**0.0234**	[0.4, 5.5]	8/10	0.1562
precision	[1.4, 11.3]	8/10	0.1562	[−0.6, 5.3]	7/10	0.2324
recall	[7.0, 15.1]	9/10	**0.0234**	[0.6, 7.5]	7/10	0.1680

D.7 Scope and outlook

Towards broader training. The reading tests span the 17-model core set and the reasoning trajectories 7 models, MedGaze is trained on SLAKE organ questions for 14 models, and report generation covers 10 models on Chest ImaGenome. Training MedGaze on report corpora with region annotations, beyond SLAKE, Qwen-VL and Gemma3, is the natural next step, and nothing here is validated for clinical use.

E Reproducibility and Engineering

Package. Discovery, steering, the split-prefill decoder, calibrated steering and the MedGaze adapter ship as one Python package, which will be released with the code upon publication. One call loads a supported checkpoint with eager attention, and further calls rank heads on images with two or more named regions, learn per-head gains, answer with the heads steered to a mask or box, and save the heads and gains. Steering is a context manager, so every call inside the block is steered, and the adapter needs no region. A new model family is one adapter subclass that states five facts, the image-token span, the grid geometry, the pixel-to-token map, the attention modules and the query-head count. Every table and figure is regenerated from the run files under `results/` by the scripts named in each file's header.

Compute and memory. All runs use `bfloat16` and eager attention on NVIDIA H200 GPUs, with Python 3.11.16, PyTorch 2.14.0 on CUDA 13.0 and Transformers 5.16.1, since fused attention kernels never materialise the attention matrix. Discovery reads one row per layer, so a forward hook keeps only the final token's row, and retained memory grows linearly rather than quadratically with sequence length, which fits the largest models on one card. Inputs are capped at 1024^2 pixels, because one layer's attention matrix for a native-resolution chest radiograph exceeds one GPU, and weights larger than one card are split over two GPUs by layer.

Implementation notes. (i) *Split prefill.* The last prompt token runs as its own steered step, because decode-time hooks do not see the prompt row that predicts the first answer token, as Appendix B.3 details. (ii) *Degeneracy filter.* Degenerate outputs count as misses and every comparison uses a size-matched control set, since steering every head produces looping text that leaks content words and can fool a judge. (iii) *Sink position.* The attention sink sits at position 0, 1 or 2 depending on the chat template and is always measured at its true position. (iv) *Neutral prompts.* Every read-out uses a neutral prompt, because a prompt that states the finding lets log-probability measure copying. (v) *Full decode.* Every DICOM file is fully decoded before use, since truncated files can decode with only a warning.

F Qualitative Examples

Individual SLAKE studies show the attention shifts and the generated text behind the aggregate results.

Expert review. An expert radiologist reviewed every qualitative example in this paper before inclusion. The review checked the anatomy and organ labels on each attention map and whether each steered sentence describes the organ or region it was steered to.

How the examples are drawn. Quotes are verbatim model output on open-licence SLAKE studies that do not appear in the main figures, except the two cases enlarged in the top row of Figure 13. Each map is the mean attention of the 100 fixation heads over image tokens, scaled to its own maximum for display, whereas every printed percentage uses the raw attention.

Organ VQA. Maps are taken at the first answer token, and the percentage is the share inside the organ's box (dashed). Below its top row, Figure 13 shows for each model the case with the largest gain in organ share among those the frozen model answers wrongly and MedGaze correctly (green), then cases correct in both whose share rises by at least 20 points (blue), excluding lesions, run-on answers and spelling-only differences.

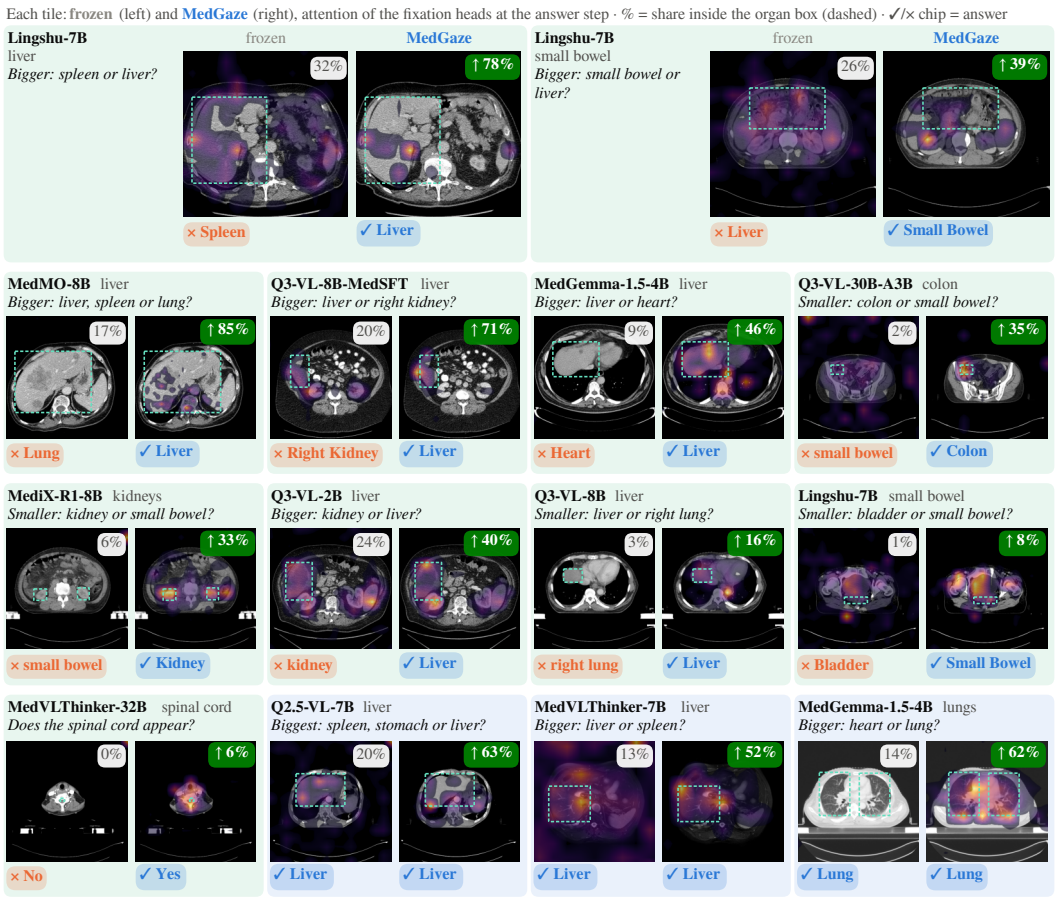


Figure 13: **MedGaze moves attention onto the organ and corrects the answer.** Top row: two `Lingshu-7B` cases, enlarged; then 12 tiles, frozen (left) against MedGaze (right), no region at test.

Anatomy-conditioned generation. Steered to an organ mask, the fixation heads put nearly all their attention inside it and the model names that organ, as Figure 14 shows on four more SLAKE studies. The frozen model is steered without any adapter, every tile uses the prompt "Describe any abnormality visible in the highlighted part of this image. If it appears normal, say so.", and the map is

averaged over the generated tokens with the percentage taken inside the organ mask. The examples are the studies whose steered first sentence names the steered organ, shown with two steered organs per study where a study has them and otherwise with the unsteered output beside the one steered organ.

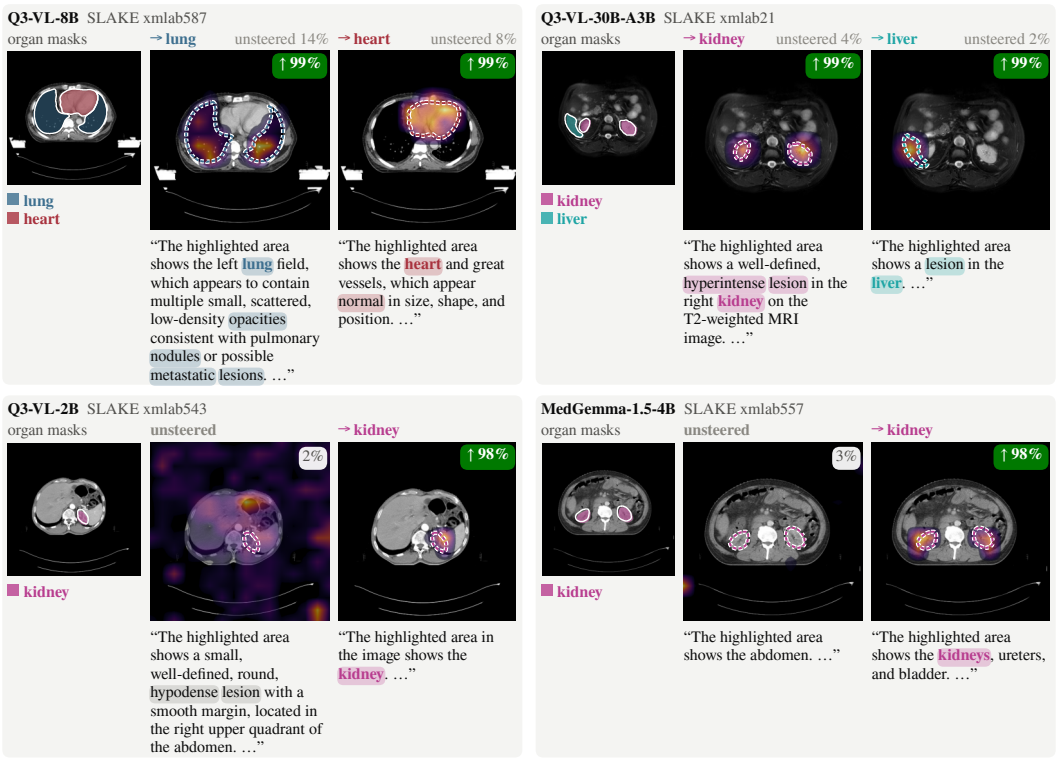


Figure 14: **Steering one organ at a time makes the model describe that organ.** Four SLAKE studies; the steered mask is dashed in its organ colour and the badge gives the share of attention inside it.

Every condition, one case per pilot. Figure 15 shows one closed organ question per pilot model under five conditions, (i) unsteered, (ii) the fixation heads steered to the named organ at $b = 4$, (iii) the same heads steered to a different organ, (iv) as many random heads steered to the named organ, and (v) MedGaze with no region. Each map is the fixation heads' attention at the first answer token, and the badge is the share inside the organ's box. Each case is an image in no other figure, with an organ rather than a lesion and all five conditions captured, and frozen-wrong cases come first with ties broken by the MedGaze share gain. Random heads leave the share at the unsteered level in every row, whereas the fixation heads aimed at the organ put nearly all their attention inside it.

How the report cards are chosen. Figure 16 shows credentialed Chest ImaGenome studies, so no reference report text appears, and the chips are the study's finding labels (green ✓ recovered, blue asserted by both, dashed missed by both). A pair is eligible when the method asserts a reference-positive finding that the baseline, the frozen model or the unsteered tour, omitted or denied, loses no finding, adds no unsupported one and stays under the token cap, which 38 MedGaze and 155 tour pairs of 1,420 satisfy.

How the card maps are made. Each map teacher-forces the shown generation (greedy agreement 93% to 100%), averages the fixation heads over the highlighted sentence and boxes the region of the recovered finding. Cards are shown when the gained attention inside the box is at least twice the box's share of the image and the in-box share rises by at least five points, and on identical tokens MedGaze puts 2 to 8 points more attention inside the box.

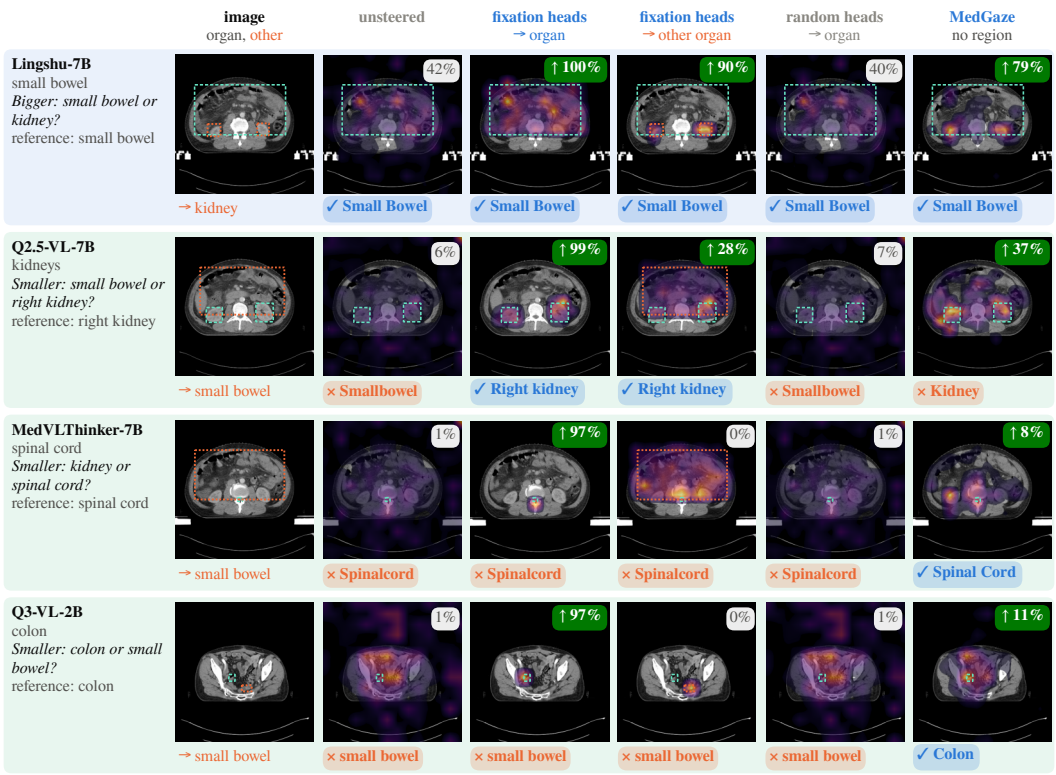


Figure 15: **Aiming the fixation heads, or MedGaze with no region, moves attention onto the organ, whereas random heads do not.** One organ question per pilot model under every condition.

Open chest X-rays and CT. Figure 17 uses SLAKE CT slices and NIH ChestX-ray14 X-rays outside the adapter's training images, each with its dataset box, and frozen and MedGaze generate from the same image, prompt, seed and budget. Maps average the heads over the phrase naming the finding, or over the whole report when one report omits it. An X-ray tile is shown when MedGaze names the finding and the frozen report does not, or puts at least two more points inside the box, and a CT tile when MedGaze adds to the lesion description or keeps it with a higher in-box share. Box labels are dataset annotations, and no diagnosis is verified.

Every example is drawn by the fixed rules above, and Appendix E lists what is needed to reproduce them.

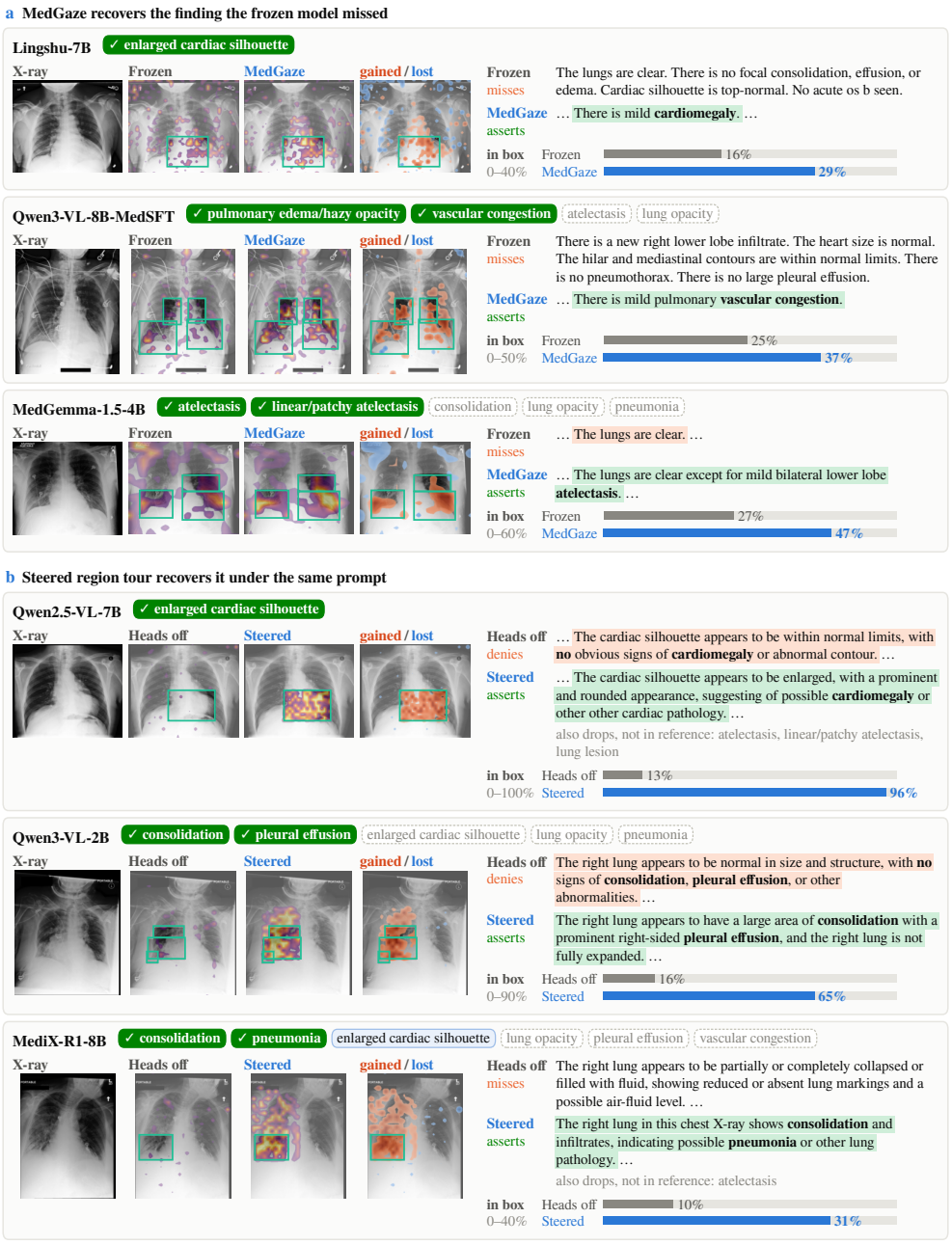


Figure 16: **Steering and MedGaze move the fixation heads onto the finding they recover.** Each card shows the X-ray and the heads' attention while the model writes the highlighted sentence, with the study's reference findings as chips and the finding's region as a box.

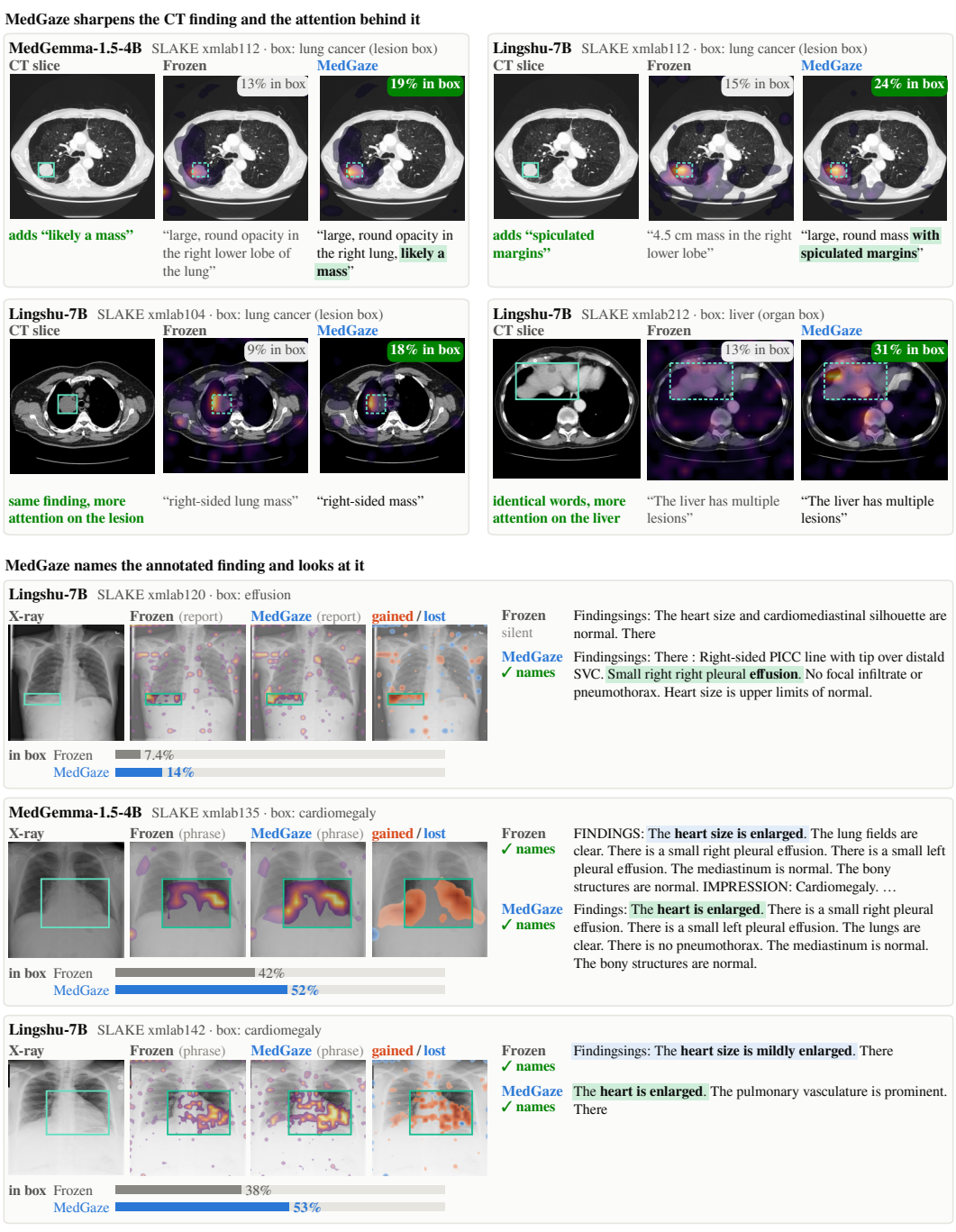


Figure 17: **MedGaze sharpens open findings, and the fixation heads' attention while writing them moves into the dataset box.** (a) SLAKE CT phrases and (b) SLAKE chest X-ray reports, frozen against MedGaze; badge: share inside the box.