LEARNING WHERE TO LOOK: REINFORCEMENT LEARNING FROM RADIOLOGIST GAZE FOR VISUALLY GROUNDED CHEST X-RAY VQA

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent medical vision-language models have shown strong capability in chest X-ray (CXR) interpretation. Supervised fine-tuning (SFT), which optimizes the answer tokens, has enabled this capability through instruction tuning, and reinforcement learning (RL) has further improved it by updating the policy with rewards on answer correctness or reasoning quality. However, these training signals supervise only the generated text rather than how the model reaches the answer from the image. Neither objective constrains where the model looks, allowing a model to answer correctly while attending to clinically irrelevant regions. We propose CheXGaze, a CXR-specialized VLM for visual question answering that learns where to look by using radiologists' eye gaze as an RL reward rather than a target to imitate. Following radiologists' global-to-focal reading process, CheXGaze generates an `<organ>` bounding box, `<gaze>` points, and an `<answer>`. Experiments on MIMIC-CXR-VQA and the CXR subsets of SLAKE and VQA-RAD demonstrate its effectiveness, and attention analysis shows closer alignment with radiologists' gaze, indicating that CheXGaze learns where to look.

1 INTRODUCTION

Medical vision-language models (VLMs) have shown strong capabilities across diverse tasks, including clinical diagnosis, visual question answering, and report generation (Li et al., 2023; Chen & Hong, 2024; Sellergren et al., 2025). Chest X-rays (CXRs) are particularly important among diverse modalities, serving as a primary tool for diagnosing and monitoring diseases (Suetens, 2017), with more than 2 billion performed annually (Akhter et al., 2023). However, CXR interpretation requires expert radiologists and approximately two-thirds of the world's population lacks access to radiology diagnostics (Mollura et al., 2010; Kesselman et al., 2016). In this context, medical VLMs specialized for CXR can offer considerable potential to support clinical diagnosis and decision-making (Chen et al., 2024c; Zambrano Chaves et al., 2025).

Regarding the potential of medical VLMs in clinical diagnosis, prior work has progressed from supervised fine-tuning (SFT)-based instruction tuning (e.g., LLaVA-Med, HuatuoGPTVision) (Li et al., 2023; Chen et al., 2024a) to reinforcement learning (RL)-based approaches for eliciting stronger reasoning capabilities (e.g., MedVLM-R1, Med-R1)(Pan et al., 2025; Lai et al., 2026). Recent CXR-specialized VLMs such as CheXOne (Zhang et al., 2026) and CX-Mind (Li et al., 2025) combine large-scale instruction tuning with RL to improve diagnostic performance while generating explicit reasoning traces.

Despite such progress, a fundamental question remains: *"Does the model actually look at the clinically relevant region?"* Although the visual features are used to generate the answer, the training signals of both SFT and RL are computed only on the generated text, providing no explicit feedback on how the answer is derived from the image. SFT methods are trained on the answer tokens, while RL rewards answer correctness or reasoning quality. Such text-centric supervision does not constrain where the model looks, so a model can answer correctly while attending to clinically irrelevant regions. This limitation is especially critical for chest X-rays, where anatomical structures are highly similar across patients and findings are often small or subtle, making visually grounded reasoning essential (Han et al., 2022; Çallı et al., 2021).

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.

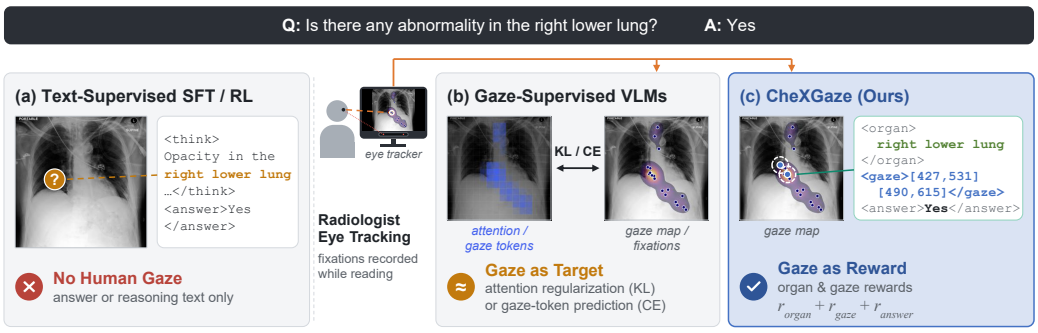

Figure 1: Comparison between prior approaches and ours.

Accordingly, several recent works have explored teaching VLMs *where to look*. MedEyes (Zhu et al., 2026) uses RL to progressively attend to relevant image regions, but relies on synthetic visual trajectories rather than actual expert behavior. Radiologists' eye gaze offers a more direct signal of clinically relevant evidence, and has been used to supervise model attention (Pani & Yang, 2026) or to train gaze tokens that predict fixated image patches (Li et al., 2026; Lee et al., 2026). However, these approaches treat gaze as a supervised target to imitate. Since radiologists can exhibit clinically valid yet different viewing patterns (Kelahan et al., 2019; Rajaraman et al., 2020), such supervision reduces expert guidance to imitating their gaze pattern.

To address this limitation, we propose CheXGaze, a CXR-specialized VLM that leverages radiologists' eye gaze as an RL reward rather than just a target to imitate. Whereas SFT trains the model to imitate the recorded gaze through losses, CheXGaze generates its own gaze points and is rewarded by the fraction of the radiologist's gaze mass they capture. Such RL-based training enables exploration and updates the policy toward capturing more gaze evidence, encouraging the model to *learn where to look* and to generalize better than SFT-based imitation (Chu et al., 2025). Following the global-to-focal reading process of radiologists, CheXGaze generates the relevant organ, two gaze points, and the answer in sequence. To the best of our knowledge, CheXGaze is the first to use real radiologists' gaze as an RL reward for medical VLMs.

Specifically, we focus on CXR visual question answering in this work. We utilize REFLACX and EGD-CXR (Bigolin Lanfredi et al., 2022; Karargyris et al., 2020), which record radiologists' gaze on MIMIC-CXR images with their dictated reports. We temporally align each dictated finding with its fixations and match it to the corresponding MIMIC-CXR-VQA (Bae et al., 2024) question, yielding 2,494 QA pairs, only 0.86% of the full training set. Given this scarcity, we alternate SFT on the full data with Group Relative Policy Optimization (GRPO) on the gaze subset with a combined reward consisting of three terms: `<organ>`, `<gaze>`, and `<answer>`. This iterative SFT–GRPO training keeps the policy from drifting toward the small subset, making it practical in real-world settings where gaze is sparse. Experimental results show that CheXGaze outperforms general-purpose, medical, and CXR-specialized VLMs on MIMIC-CXR-VQA and improves over the same backbone fine-tuned without gaze on SLAKE and VQA-RAD (Liu et al., 2021; Lau et al., 2018), which contain no gaze annotations. Further analysis shows that its attention aligns more closely with radiologists' gaze, suggesting that CheXGaze ultimately learns *where to look*.

2 Related Works

2.1 Supervised Fine-tuning for Medical VLMs

General-purpose VLMs such as LLaVA (Liu et al., 2023), Qwen2.5-VL (Bai et al., 2025), and InternVL (Chen et al., 2024b) have demonstrated strong multimodal understanding via large-scale pretraining. Building upon them, medical VLMs such as LLaVA-Med (Li et al., 2023), HuatuoGPT-Vision (Chen et al., 2024a), and MedGemma (Sellergren et al., 2025) are adapted to medical domain through supervised instruction tuning, optimizing token-level cross-entropy on medical image–text pairs. CXR-specialized models further adapt VLMs to CXR interpretation. CheXagent (Chen et al., 2024c) introduces CheXinstruct, comprising 8.5M training samples, and performs multi-task instruction tuning. LLaVA-Rad (Zambrano Chaves et al., 2025) employs a CXR-specific vision en-

coder pretrained on 697K image–report pairs, followed by an alignment stage and supervised fine-tuning on MIMIC-CXR (Johnson et al., 2024).

2.2 Reinforcement Learning for Medical VLMs

While SFT is effective for learning from target responses, it essentially imitates them (Guo et al., 2024) and can lead to memorization and limited out-of-distribution generalization. In contrast, RL optimizes the policy to maximize expected reward, allowing the model to explore beyond explicit supervision (Williams, 1992). In practice, the two are often combined, with SFT stabilizing the output format and subsequent RL improving out-of-distribution generalization through reward signals (Chu et al., 2025).

In the medical domain, MedVLM-R1 (Pan et al., 2025) applies GRPO directly to Qwen2-VL with accuracy and format rewards, using 600 medical VQA samples, and Med-R1 (Lai et al., 2026) extends this to eight imaging modalities. However, both report that generated reasoning is not necessarily faithful, can produce hallucinated rationales, and may even degrade performance. MediX-R1 (Mullappilly et al., 2026) extends RL to open-ended medical VLM tasks with LLM-based accuracy, embedding-based semantic similarity, format, and modality recognition rewards. For CXRs, CheXOne (Zhang et al., 2026) and CX-Mind (Li et al., 2025) apply GRPO after large-scale instruction tuning, and CX-Mind further introduces verifiable process rewards to reduce hallucinations. Overall, these reward designs range from outcome-level accuracy and format rewards to semantic, modality-specific, and process-level rewards, but they are all computed from the textual output and do not account for where the model looks.

2.3 Gaze Supervision

Despite the aforementioned advances in VLMs, their reasoning is generally carried out through text, which can be suboptimal for visually grounded tasks. Thus, several studies have incorporated human gaze data, which provides local evidence of where visual information should be acquired. In the general domain, Gaze-VLM (Pani & Yang, 2026) aligns VLM attention with human gaze distributions via KL-divergence regularization loss. In the medical domain, 'Thinking with gaze' (Li et al., 2026) converts radiologist gaze into discrete image–patch indices and trains the gaze tokens to predict them using cross-entropy loss. Gaze-X (Lee et al., 2026) incorporates gaze trajectories and fixation patterns into pretraining through gaze token prediction, and transfers to downstream tasks including report generation, visual grounding, and VQA.

These gaze-aware approaches incorporate gaze as supervised objectives, treating the gaze patterns as targets. Since gaze patterns vary across radiologists (Appendix A), such supervision reduces to imitating an individual reader's gaze. Instead, we use gaze as an RL reward, so that the model explores its own gaze points and updates its policy toward those capturing more of the radiologist's gaze mass, encouraging generalization beyond imitation (Chu et al., 2025). To the best of our knowledge, we are the first to use radiologist gaze as an RL signal in a CXR-specialized VLM.

3 Method

Figure 2 illustrates the overall framework of CheXGaze for CXR-VQA. In this section, we (1) formulate the task, (2) describe the data preprocessing procedure, and (3) detail the training pipeline.

3.1 Task Formulation

Given a chest X-ray image I and question Q, our goal is to train a generative VLM π_θ to generate a response $Y \sim \pi_\theta(\cdot \mid I, Q)$. Motivated by the global-to-focal visual search pattern of radiologists (Pietrzyk et al., 2008), the model is prompted to generate three tagged components in sequence:

`<organ>... </organ>` $\rightarrow$ `<gaze>...</gaze>` $\rightarrow$ `<answer>...</answer>`,

which contain the relevant anatomical regions with their bounding boxes $[x_1, y_1, x_2, y_2]$, two points $Z = (z_1, z_2)$ where the model should look, and the answer. All coordinates are normalized as integers in $[0, 1000)$ with respect to the image dimensions, with the top-left corner as $(0, 0)$. Note that radiologist gaze data is used only during training. We detail the full prompt template in Appendix B.

3.2 Data Preprocessing

The data preprocessing consists of three steps: transcript span alignment, organ target construction, and gaze target construction.

Transcript Span Alignment We align the VQA dataset MIMIC-CXR-VQA with the gaze datasets, EGD-CXR and REFLACX, that share identical DICOM IDs. Both gaze datasets provide the radiologist's full gaze trajectory with time-aligned dictation transcripts. Since each transcript covers multiple findings across the entire CXR, we use two complementary matching strategies to align each QA pair with the relevant content. We first apply keyword matching, and as many QA pairs do not have one-to-one lexical matching, we additionally use an LLM (Qwen3.6-27B) to match the span semantically. The LLM is used only to select which recorded transcript span corresponds to the question and does not produce additional gaze coordinates. Gaze fixations recorded within the matched time intervals are then retained. Table 1 reports the number of gaze-annotated QA pairs per split. Only the training pairs ($\mathcal{D}_{\text{gaze}}$) are used for training, and gaze is never used at inference. As only a few test images have gaze data, we use the validation pairs instead for the analysis in Sec. 4.3. Further details are provided in Appendix C.

Split	# QA pair	# Gaze	Coverage
Train	290,031	2,494	0.86%
Valid	73,567	723	0.98%
Test	13,793	–	–

Table 1: Gaze-annotated QA pairs per split.

Organ Target Construction To obtain the organ targets O^*, we map each question to anatomical regions from Chest ImaGenome (Wu et al., 2021). For objects mentioned in the questions, we use their bounding box directly, and for those not mentioned, we map to the highest prior related to the attribute. We do not label the bounding box of the answer to prevent leakage, and for questions without an anatomical referent (e.g., gender, plane) no regions are assigned. Specific details are in Appendix D.

Gaze Target Construction As the eye-tracking data derived by the radiologists are noisy and reader-dependent, it is not feasible to directly use all fixations as target gaze points. The median number of fixations (i.e., time interval where gaze is stabilized at a point) is 16 per QA pair. Therefore, we convert fixations into a duration-weighted fixation heatmap (Bigolin Lanfredi et al., 2022), a spatial density map that reflects where and for how long the reader looked. Each fixation $i \in F_q$ at location g_i with dwell time w_i is rendered as a Gaussian of 1° visual angle, and the result is normalized into a probability map over grid cells $u \in \mathcal{U}$:

$$H_q(u) = \sum_{i \in F_q} w_i \exp\left(-\frac{\|c(u) - g_i\|^2}{2\sigma_i^2}\right), \qquad h_q = \frac{H_q}{\sum_{u \in \mathcal{U}} H_q(u)}, \tag{1}$$

where $c(u)$ is the center of cell u and $\sigma_i = \sigma_{\text{deg}}\, \rho_i$ converts 1° to pixels via the pixels-per-degree ρ_i. The 1° scale, standard in saliency evaluation, accounts for the size of the human fovea (i.e., how much of an image a human has in focus during a fixation) and the measurement error (Bylinskii et al., 2018). This results in a median of two modes, so we use two gaze points for both targets and the predictions.

However, the fixation heatmap does not yet serve as a reward target. This is because CXRs have a center-bias issue. Since they share a highly consistent anatomical layout, radiologists fixate on similar regions regardless of the question, and h_q is thus dominated by the population prior p (the average fixation heatmap over training pairs). Lanfredi et al. (2021) report that an anatomy-only lung segmentation baseline achieved a normalized cross-correlation (NCC) of 0.637 with radiologist gaze, comparable to the inter-observer agreement of 0.632. In our framework, a constant policy that ignores the image and the question and simply outputs the two peaks of the population prior gets 32% of the attainable reward. To prevent such shortcut, we subtract the least-squares projection of h_q onto the population prior p to retain the cells where the *reader looked more*:

$$\tilde{h}_q = \max\big(0,\ h_q - \beta_q \cdot p\big), \qquad \beta_q = \frac{\langle h_q, p \rangle}{\langle p, p \rangle}. \tag{2}$$

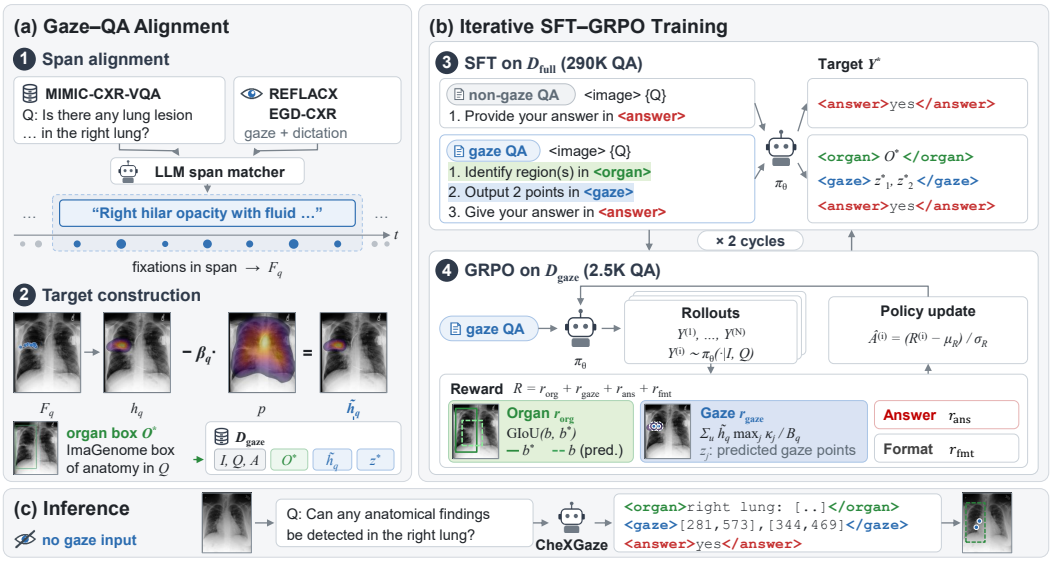

Figure 2: Overall framework of CheXGaze.

Note that the optimal gaze points on h_q still attain a median of 99.5% of the attainable reward on $\tilde{h}_q$, while the reward of the constant policy drops from 32% to 17%. See Appendix E for details.

3.3 Overall Training Pipeline

Our training pipeline alternates supervised fine-tuning (SFT) on the full VQA corpus $\mathcal{D}_{\text{full}}$ with gaze-guided GRPO on the gaze-annotated subset $\mathcal{D}_{\text{gaze}}$ (Table 1). Specifically, we perform two cycles of SFT followed by GRPO,

$$\text{SFT}(\mathcal{D}_{\text{full}}) \rightarrow \text{GRPO}(\mathcal{D}_{\text{gaze}}) \rightarrow \text{SFT}(\mathcal{D}_{\text{full}}) \rightarrow \text{GRPO}(\mathcal{D}_{\text{gaze}}).$$

Each stage serves complementary roles: SFT fits the model to the answer distribution and structured output format, while GRPO further optimizes the policy using answer, organ localization, and gaze rewards. Repeating this cycle prevents the policy from drifting toward the small gaze subset.

3.3.1 Supervised Fine-Tuning

We perform SFT on a joint corpus containing both gaze-annotated and non-gaze VQA pairs. The supervision format depends on whether gaze annotation is available. For non-gaze examples, the model is trained to directly generate the answer, `<answer>...</answer>`, whereas gaze-annotated examples follow the structured format of Sec. 3.1, `<organ>...</organ><gaze>...</gaze><answer>...</answer>`. For gaze-annotated samples, the `<organ>` target is O^* and the `<gaze>` target is the two point capturing the most mass of $\tilde{h}_q$. We optimize all target tokens using standard autoregressive cross-entropy. This joint SFT enables the model to generate structured outputs while learning from the full VQA corpus, thereby providing an initial policy for subsequent gaze-guided GRPO.

3.3.2 Gaze-Guided GRPO

Following SFT, we perform GRPO to optimize the policy π_θ on $\mathcal{D}_{\text{gaze}}$ (Shao et al., 2024). For each image-question pair (I, Q), we sample a group of N responses $\{Y^{(i)}\}_{i=1}^{N}$ and assign each response a scalar reward $R^{(i)} = R(Y^{(i)})$. The group-relative advantage is computed as:

$$\hat{A}^{(i)} = \frac{R^{(i)} - \mu_R}{\sigma_R + \epsilon},$$

where μ_R and σ_R denote the mean and standard deviation of rewards within the group. The policy is then updated using the standard clipped GRPO objective with KL regularization against the reference policy (the previous SFT model).

Composite Reward We optimize GRPO using a composite reward consisting of answer correctness, output format, organ localization, and gaze grounding:

$$R(Y) = w_{\text{ans}} r_{\text{ans}} + w_{\text{fmt}} r_{\text{fmt}} + w_{\text{org}} r_{\text{org}} + w_{\text{gaze}} r_{\text{gaze}},$$

where all reward terms are normalized to $[0, 1]$ and we set all weights to 1 in our experiments. The answer reward combines accuracy and micro-F1, our evaluation metrics,

$$r_{\text{ans}} = 0.5\,\text{Acc} + 0.5\,\text{F1},$$

since many MIMIC-CXR-VQA questions require a list of items (Sec. 4.1). The format reward r_{fmt} is 1 when the response follows the required `<organ>` → `<gaze>` → `<answer>` structure and 0 otherwise. For organ localization, r_{org} is the mean rescaled GIoU over $|O^*|$ target boxes under greedy one-to-one matching on box coordinates only, scoring only the first $|O^*|$ predicted boxes. For QA pairs without any target organ box, r_{org} is 1 only if no box is predicted.

Gaze Reward Given the two predicted gaze points $Z = (z_1, z_2)$, we place a soft spatial kernel κ_j around each point on the gaze-target grid $\mathcal{U}$. We combine the two kernels by their pointwise maximum to form the predicted claim field

$$\Phi_Y(u) = \max_{j \in \{1,2\}} \kappa_j(u), \qquad u \in \mathcal{U}.$$

Let $\tilde{h}_q$ denote the corrected gaze target for QA pair q constructed in Sec. 3.2. The mass captured by prediction Y is

$$M_q(Y) = \sum_{u \in \mathcal{U}} \tilde{h}_q(u) \Phi_Y(u).$$

We normalize this quantity by the maximum attainable mass B_q for the same target:

$$r_{\text{gaze}}(Y) = \frac{M_q(Y)}{B_q} \in [0, 1].$$

Thus, r_{gaze} measures the fraction of the available gaze mass captured by the predicted points, rather than their geometric distance from a single target coordinate. We use a soft kernel because a hard spatial threshold assigns the same zero reward to all predictions outside the tolerance region, reducing the reward variation required by GRPO. The two kernels are combined by a maximum rather than a sum so that placing both gaze points at the same location does not receive duplicate credit. Similarly, to prevent reward hacking by emitting many points, only the first two predicted coordinates are scored and any additional points are ignored, as in r_{org}. Finally, the normalization term B_q accounts for variation in the attainable gaze mass across QA pairs, preventing the effective contribution of r_{gaze} to the composite reward from varying with the target distribution.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate CheXGaze on three medical VQA benchmarks: MIMIC-CXR-VQA, SLAKE, and VQA-RAD. Since SLAKE and VQA-RAD cover multiple imaging modalities, we only use their chest X-ray subsets, denoted as SLAKE-CXR and VQA-RAD-CXR, which contain 1,423/361 and 1,088/174 train/test QA pairs, respectively.

Baselines We compare against a broad range of VLMs in three categories: (i) ***General VLMs***: Qwen2.5-VL-3B/7B Bai et al. (2025), InternVL3.5-8B Wang et al. (2025), Gemma3-4B Team et al. (2025), and LLaVA-1.5-7B Liu et al. (2023). (ii) ***Medical VLMs***: MedGemma-4B Sellergren et al. (2025), HuatuoGPT-Vision-7B Chen et al. (2024a), LLaVA-Med v1.5 Li et al. (2023), MedVLM-R1 Pan et al. (2025), and MediX-R1 Mullappilly et al. (2026). (iii) ***CXR-specialized VLMs***: CX-Mind Li et al. (2025), CheXOne Zhang et al. (2026), LLaVA-Rad Zambrano Chaves et al. (2025), and CheXagent Chen et al. (2024c).

Table 2: Performance comparison on MIMIC-CXR-VQA, SLAKE-CXR, and VQA-RAD-CXR. For MIMIC-CXR-VQA, we report accuracy for 'verify' and F1 for 'choose'/'query' questions. The 'Acc.' and 'F1' denote overall metrics, computed over all three question types and over 'choose' and 'query', respectively. For SLAKE-CXR and VQA-RAD-CXR, we report accuracy and recall for closed- and open-ended questions, respectively. Fine-tuned VLMs are averaged over 3 runs. **Bold** and underlined denote the best and second-best.

		MIMIC-CXR-VQA					SLAKE-CXR		VQA-RAD-CXR	
Model	**#Param**	**Acc.**	**F1**	**Verify**	**Choose**	**Query**	**Closed**	**Open**	**Closed**	**Open**
General VLMs (Zero-Shot)										
Qwen2.5-VL-3B	3B	40.83	27.82	56.43	33.42	26.22	61.40	36.22	62.93	22.36
Qwen2.5-VL-7B	7B	42.90	31.84	56.80	34.17	31.18	78.07	41.67	68.97	31.67
InternVL3.5-8B	8B	43.80	31.42	62.09	44.65	27.64	73.68	55.57	69.83	38.25
Gemma3-4B	4B	30.08	15.98	50.99	37.47	9.84	82.46	41.19	45.69	23.22
LLaVA-1.5-7B	7B	41.87	32.19	53.15	35.22	31.32	59.65	35.89	50.86	10.06
Medical VLMs (Zero-Shot)										
MedGemma-4B	4B	45.25	27.65	68.73	49.98	21.26	**94.74**	48.46	76.72	32.17
HuatuoGPT-Vision-7B	7B	46.04	34.37	60.63	39.59	32.88	70.18	42.88	67.24	37.70
LLaVA-Med	7B	25.10	1.06	49.78	3.14	0.47	69.30	45.91	56.03	33.25
MedVLM-R1	2B	43.25	32.07	56.00	36.34	30.85	66.67	49.64	66.38	29.75
MediX-R1	8B	40.73	26.72	58.75	38.64	23.31	91.23	68.19	73.28	40.33
CXR-Specialized VLMs (Zero-Shot)										
CX-Mind	7B	42.13	26.41	64.81	56.92	17.69	74.56	31.22	76.72	27.13
CheXOne	3B	55.82	40.91	**79.48**	54.67	36.98	76.32	43.31	74.14	33.71
LLaVA-Rad	7B	33.79	3.80	63.73	6.82	2.94	67.54	23.90	63.79	39.05
CheXagent	8B	52.72	44.97	72.28	45.94	44.69	68.42	29.92	68.10	27.82
Fine-Tuned VLMs										
LLaVA-Med-7B	7B	50.68	56.05	53.62	58.37	55.39	84.50	82.09	80.17	61.87
Qwen2.5-VL-7B	7B	65.46	64.45	77.47	65.85	64.05	84.21	82.27	**85.34**	57.03
Gaze-VLM	7B	65.43	64.19	77.47	65.03	63.95	77.20	73.88	81.32	55.46
CheXGaze (Ours)	**7B**	**66.18**	**65.24**	78.87	**65.92**	**65.05**	84.79	**85.03**	**85.34**	**65.73**

General VLMs serve as reference zero-shot baselines, while the medical and CXR-specialized models assess transfer from domain-specific pretraining. Note that several models are already exposed to our benchmark data, not reflecting a strict zero-shot setting. CX-Mind, CheXOne, and CheXagent are trained on MIMIC-CXR-VQA, MedGemma, MediX-R1, and LLaVA-Rad on MIMIC-CXR images and reports, and MedGemma, MedVLM-R1, MediX-R1, and CheXOne on the training splits of SLAKE and VQA-RAD (see Table A4 in Appendix F). For a controlled comparison, we fine-tune LLaVA-Med, Qwen2.5-VL-7B, and Gaze-VLM Pani & Yang (2026) on the same data as ours, reproducing Gaze-VLM on our Qwen2.5-VL-7B backbone with its official implementation. To evaluate whether the learned gaze transfers to datasets without gaze annotations, we adapt Gaze-VLM and CheXGaze to SLAKE-CXR and VQA-RAD-CXR from their MIMIC-CXR-VQA fine-tuned weights. LLaVA-Med and Qwen2.5-VL-7B adapt better from the base model than the model trained on MIMIC-CXR-VQA, so we use the stronger initialization. We report the average over three runs for the fine-tuned models. Full results with standard deviation are reported in Appendix G.

Evaluation Metrics MIMIC-CXR-VQA consists of *'verify'* (yes/no), *'choose'* (selecting among given candidates), and *'query'* (listing all items) questions. We report accuracy for 'verify' and micro-F1 for 'choose' and 'query'; overall Acc. and F1 are computed over all questions and over choose and query, respectively. For SLAKE-CXR and VQA-RAD-CXR, we report accuracy for closed-ended and recall for open-ended questions, following LLaVA-Med Li et al. (2023).

Implementation Details. We use Qwen2.5-VL-7B as the base model and apply full fine-tuning on 4 NVIDIA H100 GPUs. The learning rate is 1×10^{-5} for SFT and 1×10^{-6} for GRPO. For GRPO, we sample $N{=}4$ completions per prompt and set the KL coefficient to 0.04. The maximum completion length is 640 tokens, and the maximum image resolution is 401,408 pixels.

4.2 Main Result and Ablation Studies

Table 2 presents a performance comparison between CheXGaze and the baselines. Zero-shot VLMs show limited performance despite prior exposure to some of our benchmarks, confirming the necessity of task-specific adaptation. Among fine-tuned VLMs, CheXGaze achieves the best performance

on all three benchmarks. Although zero-shot MedGemma and MediX-R1 attain higher closed-ended accuracy on SLAKE-CXR, they fall behind all fine-tuned models on open-ended questions. Notably, compared with Qwen2.5-VL-7B fine-tuned on the same data without gaze, CheXGaze improves performance even though gaze is available for only 0.86% of the training data, whereas Gaze-VLM yields no improvement. This indicates that using gaze as an RL reward is more effective than aligning attention with gaze during supervised fine-tuning. Furthermore, the performance on SLAKE-CXR and VQA-RAD-CXR suggests that the visual grounding learned from gaze supervision on MIMIC-CXR-VQA is generalizable to datasets without gaze annotations.

Organ & Gaze Tag Ablation To examine the contribution of the `<organ>` and `<gaze>` tags, we conduct an ablation study removing each component and report the results in Table 3. Due to space constraints, we report overall accuracy and F1 for MIMIC-CXR-VQA, and the sample-weighted average of closed-ended accuracy and open-ended recall over closed- and open-ended questions for SLAKE and VQA-RAD. Since the CXR subset of SLAKE and VQA-RAD is relatively small, we report the average over three runs. The results show that using the organ tag alone provides no consistent improvement over the model without tags, whereas the gaze tag improves performance on most metrics. Furthermore, using both tags yields the best overall results, suggesting that gaze provides useful visual supervision and anatomical information offers additional complementary guidance.

Table 3: Ablation on `<organ>` and `<gaze>` tags.

Organ	Gaze	MIMIC-CXR-VQA		SLAKE	VQA-RAD
		Acc.	F1		
✗	✗	65.59	64.68	81.27	73.42
✓	✗	64.71	63.25	83.21	72.73
✗	✓	66.05	64.35	83.46	75.74
✓	✓	**66.10**	**64.82**	**84.95**	**78.80**

Step-wise Evaluation To validate our iterative SFT–RL training pipeline, we perform a step-wise evaluation on MIMIC-CXR-VQA (Table 4). Each '+' row stands for continuing training from the preceding row. Our iterative training pipeline progressively improves performance across stages and achieves the best overall results after the second GRPO stage. In contrast, applying GRPO for two epochs after SFT does not lead to performance improvement, even reducing accuracy. We assume this is due to the limited size of the gaze-annotated subset (0.86% of the full training data), making the model biased toward the limited subset distribution. Alternating full-data SFT with gaze-guided GRPO helps mitigate this effect, supporting iterative SFT–RL training in real-world settings with sparse gaze supervision.

Table 4: Step-wise evaluation on MIMIC-CXR-VQA.

Cyclic	Stage	Acc.	F1	Verify	Choose	Query
✓	SFT_{1ep}	63.68	61.37	77.61	61.39	61.37
	+ $GRPO_{1ep}$	64.18	62.67	77.42	63.73	62.37
	+ SFT_{1ep}	65.40	63.74	78.53	63.51	63.81
	+ $GRPO_{1ep}$	**66.10**	**64.82**	**78.75**	66.36	**64.38**
✗	SFT_{2ep}	65.37	64.32	78.01	65.81	63.90
	+ $GRPO_{2ep}$	64.92	64.38	77.60	**66.80**	63.69

Fine-tuning on non-gaze datasets Many medical VQA datasets do not include eye-gaze annotations. Therefore, we examine whether the organ and gaze grounding learned on MIMIC-CXR-VQA is generalizable to such datasets. Specifically, we compare three fine-tuning variants: (i) *Direct FT*, which fine-tunes the model to generate only the final answer; (ii) *Predicted traces*, which uses `<organ>` and `<gaze>` traces generated by the pretrained model as additional input; and (iii) *Corrupted traces*, which uses corrupted traces as input instead. Table 5 shows that 'Direct FT' outperforms both baselines on SLAKE-CXR and competitive on VQA-RAD-CXR, demonstrating that our MIMIC-CXR-VQA fine-tuned model serves as a good initialization for other CXR

Table 5: Fine-tuning on SLAKE-CXR and VQA-RAD-CXR without gaze annotations (mean ± std over three runs).

Method	SLAKE-CXR		VQA-RAD-CXR	
	Closed	Open	Closed	Open
Baselines				
LLaVA-Med-7B	84.50±1.34	82.09±0.21	80.17±0.86	61.87±3.37
Qwen2.5-VL-7B	84.21±1.52	82.27±1.08	**85.34**±0.87	57.03±1.26
Ours (init. from MIMIC-fine-tuned)				
Direct FT	**85.96**±0.88	83.65±0.66	83.33±1.31	60.92±1.00
+ Predicted Trace	84.79±1.01	**85.03**±0.29	**85.34**±1.50	**65.73**±2.47
+ Corrupted Trace	84.21±0.88	82.70±0.48	81.90±1.73	57.95±1.16

datasets. Predicted traces further improve over Direct FT on most metrics, whereas corrupted traces consistently degrade performance. This confirms that the gain stems from the learned organ and gaze grounding rather than from the reasoning format itself.

4.3 Does CheXGaze Look Where Radiologists Look?

Occlusion Analysis To assess whether CheXGaze trained with gaze rewards actually relies on radiologist-fixated regions, we perform an occlusion analysis on the 723 gaze-annotated validation pairs. We occlude the top 2.5%, 5%, and 10% of the fixation heatmap h_q and measure the resulting performance drop Δ. The more the model relies on these regions, the higher the drop. Table 6 compares CheXGaze with a non-gaze baseline using the same Qwen2.5-VL-7B backbone. Since two models may rely on the image to different extents, we evaluate them on a fully blank image and compute the resulting drop (Full−Blank) as the image contribution.Rel. $\Delta = \Delta/(\text{Full} - \text{Blank})$ measures the fraction of the image contribution removed by region occlusion. CheXGaze consistently shows a larger drop than the baseline, suggesting that it relies more heavily on visual evidence from radiologist-fixated regions.

Table 6: Performance under fixated-region occlusion.

		Acc.			F1		
Image	Model	Score	Δ	Rel. Δ	Score	Δ	Rel. Δ
Full	w/o Gaze	62.93	–	–	60.10	–	–
	Ours	66.67	–	–	65.89	–	–
Blank	w/o Gaze	35.68	27.25	–	30.56	29.54	–
	Ours	38.17	28.49	–	40.95	24.94	–
ROI occluded							
2.5%	w/o Gaze	60.72	2.21	8.12	59.86	0.24	0.81
	Ours	62.52	**4.15**	**14.56**	60.30	**5.58**	**22.39**
5%	w/o Gaze	58.64	4.29	15.74	56.23	3.87	13.10
	Ours	58.64	**8.02**	**28.16**	55.25	**10.64**	**42.66**
10%	w/o Gaze	54.91	8.02	29.44	55.95	4.15	14.04
	Ours	58.23	**8.44**	**29.61**	54.03	**11.85**	**47.53**

Attention Map Analysis We further analyze and visualize attention maps to examine whether gaze supervision actually changes 'where the model looks'. We compare the model's attention map with the radiologist fixation map and compute the ROI ratio and the cosine similarity. The ROI ratio divides the mean attention inside the ROI (top 5% of the fixation map) by the mean over the whole image, 1 indicating chance. Compared with the baseline without gaze supervision, our model raises the ROI ratio from 1.22 to 1.58 and the cosine similarity from 0.22 to 0.27, shifting attention toward radiologist-attended regions. Figure 3 shows a representative example. The baseline's attention is scattered across the image and yields an incorrect answer, whereas our model focuses on the lower right lung and predicts correctly.

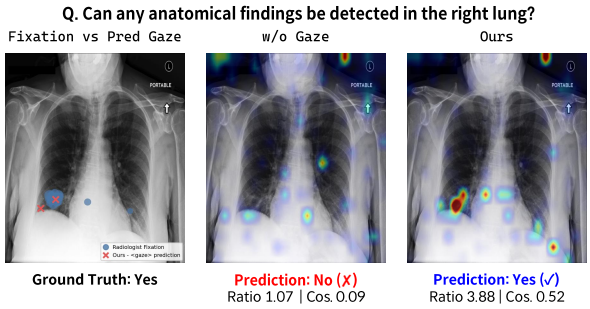

Figure 3: Attention map visualization.

5 Conclusion and Limitation

In this paper, we propose CheXGaze, a CXR-specialized VLM that uses radiologists' eye gaze as an RL reward. Our model builds upon Qwen2.5-VL, trained with iterative SFT–GRPO, sequentially generating the relevant organ, gaze points, and answer. Results on MIMIC-CXR-VQA show that CheXGaze improves over the same backbone fine-tuned without gaze, even though gaze covers only 0.86% of the training data. Evaluation on SLAKE and VQA-RAD further indicates its generalization ability to datasets without gaze. Overall, our work demonstrates the effectiveness of using gaze as a reward for visually grounded CXR-VQA.

While our framework effectively leverages radiologist gaze as a reward signal, the available gaze supervision is sparse and may include question-irrelevant fixations. This noise partly arises from gaze–speech asynchrony in transcript alignment, where radiologists may describe previously viewed regions while fixating elsewhere. Also, our current task is limited to VQA, and extending it to report generation is a promising direction for future work.

6 AI USE STATEMENT

In this work, we used generative AI tools to polish the manuscript and to assist with figure generation, outlining the panel layout. We also used to conduct exploratory data analysis (EDA) on our gaze data, which helped us to identify that a population prior dominates the gaze map. This motivated us to the population prior removal described in Section 3.2. The method design and conclusions of the experimental results were decided and verified by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Yasmeena Akhter, Richa Singh, and Mayank Vatsa. Ai-based radiodiagnosis using chest x-rays: A review. *Frontiers in big data*, 6:1120989, 2023.

Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei JI, Eric Chang, Tackeun Kim, and Edward Choi. MIMIC-Ext-MIMIC-CXR-VQA: A Complex, Diverse, And Large-Scale Visual Question Answering Dataset for Chest X-ray Images. *PhysioNet*, July 2024. doi: 10.13026/deqx-d943. URL https://doi.org/10.13026/deqx-d943. Version 1.0.0.

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

Ricardo Bigolin Lanfredi, Mingyuan Zhang, William F Auffermann, Jessica Chan, Phuong-Anh T Duong, Vivek Srikumar, Trafton Drew, Joyce D Schroeder, and Tolga Tasdizen. Reflacx, a dataset of reports and eye-tracking data for localization of abnormalities in chest x-rays. *Scientific data*, 9(1):350, 2022.

Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence*, 41(3):740–757, 2018.

Erdi Çallı, Ecem Sogancioglu, Bram Van Ginneken, Kicky G Van Leeuwen, and Keelin Murphy. Deep learning for chest x-ray analysis: A survey. *Medical image analysis*, 72:102125, 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, 2024a.

Qiuhui Chen and Yi Hong. Medblip: Bootstrapping language-image pretraining from 3d medical images and texts. In *Asian Conference on Computer Vision*, pp. 98–113. Springer, 2024.

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

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, et al. Chexagent: Towards a foundation model for chest x-ray interpretation. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*, 2024c.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *Forty-second International Conference on Machine Learning*, 2025.

Geyang Guo, Ranchi Zhao, Tianyi Tang, Xin Zhao, and Ji-Rong Wen. Beyond imitation: Leveraging fine-grained quality signals for alignment. In *International Conference on Learning Representations*, volume 2024, pp. 21340–21355, 2024.

Yan Han, Chongyan Chen, Ahmed Tewfik, Benjamin Glicksberg, Ying Ding, Yifan Peng, and Zhangyang Wang. Knowledge-augmented contrastive learning for abnormality classification and localization in chest x-rays with radiomics using a feedback loop. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1789–1798. IEEE, 2022.

Alistair Johnson, Tom Pollard, Roger Mark, Seth Berkowitz, and Steven Horng. MIMIC-CXR Database. *PhysioNet*, July 2024. doi: 10.13026/4jqj-jw95. URL https://doi.org/10.13026/4jqj-jw95. Version 2.1.0.

Alexandros Karargyris, Satyananda Kashyap, Ismini Lourentzou, Joy Wu, Matthew Tong, Arjun Sharma, Shafiq Abedin, David Beymer, Vandana Mukherjee, Elizabeth Krupinski, and Mehdi Moradi. Eye Gaze Data for Chest X-rays. *PhysioNet*, September 2020. doi: 10.13026/qfdz-zr67. URL https://doi.org/10.13026/qfdz-zr67. Version 1.0.0.

Linda C Kelahan, Allan Fong, Joseph Blumenthal, Swaminathan Kandaswamy, Raj M Ratwani, and Ross W Filice. The radiologist's gaze: mapping three-dimensional visual search in computed tomography of the abdomen and pelvis. *Journal of Digital Imaging*, 32(2):234–240, 2019.

Andrew Kesselman, Garshasb Soroosh, Daniel J Mollura, Geraldine Abbey-Mensah, James Borgstede, Dorothy Bulas, George Carberry, Danielle Canter, Farhad Ebrahim, Joanna Escalon, et al. 2015 rad-aid conference on international radiology for developing countries: the evolving global radiology landscape. *Journal of the American College of Radiology*, 13(9):1139–1144, 2016.

Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, Yuheng Li, Konstantinos Psounis, and Xiaofeng Yang. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *IEEE transactions on medical imaging*, 2026.

Ricardo Bigolin Lanfredi, Ambuj Arora, Trafton Drew, Joyce D Schroeder, and Tolga Tasdizen. Comparing radiologists' gaze and saliency maps generated by interpretability methods for chest x-rays. *arXiv preprint arXiv:2112.11716*, 2021.

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.

Kinhei Lee, Peiyuan Jing, Zhenxuan Zhang, Yue Yang, Tao Wang, Dominic C Marshall, Yingying Fang, and Guang Yang. Seeing through experts' eyes: a foundational vision-language model trained on radiologists' gaze and reasoning. *npj Artificial Intelligence*, 2026.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in neural information processing systems*, 36:28541–28564, 2023.

Wenjie Li, Yujie Zhang, Haoran Sun, Yueqi Li, Fanrui Zhang, Mengzhe Xu, Victoria Borja Clausich, Sade Mellin, Renhao Yang, Chenrun Wang, et al. Cx-mind: a pioneering multimodal large language model for interleaved reasoning in chest x-ray via curriculum-guided reinforcement learning. *Information Fusion*, pp. 104027, 2025.

Yiwei Li, Zihao Wu, Yanjun Lv, Hanqi Jiang, Weihang You, Zhengliang Liu, Dajiang Zhu, Xiang Li, Quanzheng Li, Tianming Liu, et al. Thinking with gaze: Sequential eye-tracking as visual reasoning supervision for medical vlms. *arXiv preprint arXiv:2603.06697*, 2026.

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.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Daniel J Mollura, Ezana M Azene, Anna Starikovsky, Aduke Thelwell, Sarah Iosifescu, Cary Kimble, Ann Polin, Brian S Garra, Kristen K DeStigter, Brad Short, et al. White paper report of the rad-aid conference on international radiology for developing countries: identifying challenges, opportunities, and strategies for imaging services in the developing world. *Journal of the American College of Radiology*, 7(7):495–500, 2010.

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.

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.

Anupam Pani and Yanchao Yang. Gaze-vlm: Bridging gaze and vlms through attention regularization for egocentric understanding. *Advances in Neural Information Processing Systems*, 38: 163544–163577, 2026.

Mariusz W Pietrzyk, David J Manning, Alan Dix, and Tim Donovan. Relations between physical properties of local and global image-based elements and the performance of human observers in lung nodule detection. In *Medical Imaging 2008: Image Perception, Observer Performance, and Technology Assessment*, volume 6917, pp. 282–292. SPIE, 2008.

Sivaramakrishnan Rajaraman, Sudhir Sornapudi, Philip O Alderson, Les R Folio, and Sameer K Antani. Analyzing inter-reader variability affecting deep ensemble learning for covid-19 detection in chest radiographs. *PloS one*, 15(11):e0242301, 2020.

Alessandro Rinaldo and Larry Wasserman. Generalized density clustering. 2010.

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.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Paul Suetens. *Fundamentals of medical imaging*. Cambridge university press, 2017.

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.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Joy Wu, Nkechinyere Agu, Ismini Lourentzou, Arjun Sharma, Joseph Paguio, Jasper Seth Yao, Edward Christopher Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, et al. Chest imagenome dataset. *Physio Net*, 2021.

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, et al. A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings. *Nature Communications*, 16(1):3108, 2025.

Yabin Zhang, Chong Wang, Yunhe Gao, Jiaming Liu, Maya Varma, Justin Xu, Sophie Ostmeier, Jin Long, Sergios Gatidis, Seena Dehkharghani, et al. A reasoning-enabled vision-language foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2604.00493*, 2026.

Chunzheng Zhu, Yangfang Lin, Shen Chen, Yijun Wang, and Jianxin Lin. Medeyes: Learning dynamic visual focus for medical progressive diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13916–13924, 2026.

A Variation in Radiologist Gaze

REFLACX (Bigolin Lanfredi et al., 2022) was constructed over three phases. In phases 1 and 2, radiologists read a shared set of CXRs, whereas in phase 3, each image is read by a single radiologist. We use phase 3 for our main experiments because it contains substantially more gaze-annotated images. To examine inter-reader variation, we utilize the phase 1 and 2 dataset. As shown in Figure A1, radiologists viewing the same image differ in fixation locations, scan order, and coverage density.

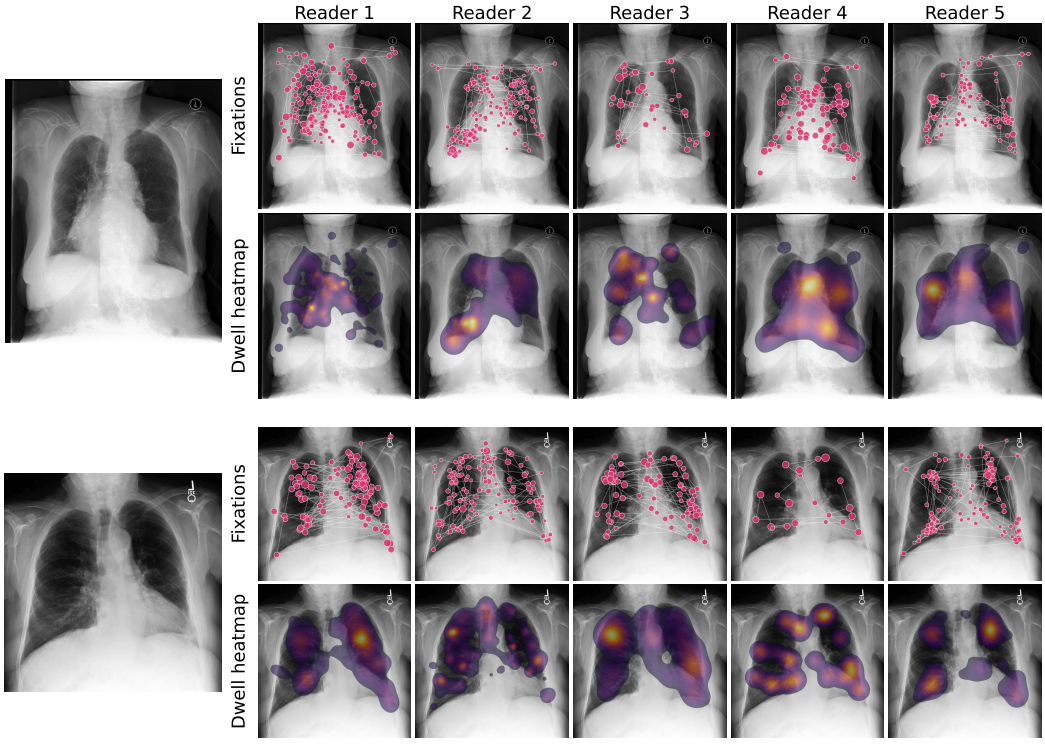

Figure A1: Gaze of five radiologists reading the same CXR from REFLACX. The top row shows fixations, with marker size proportional to fixation duration and lines showing its trajectory. The bottom row shows fixation heatmaps.

B Prompt Template

The non-gaze QA prompt in Figure A2 is used for SFT on the samples in $\mathcal{D}_{\text{full}}$ that are not in $\mathcal{D}_{\text{gaze}}$, and the gaze-annotated QA prompt in Figure A3 is used for SFT on $\mathcal{D}_{\text{gaze}}$, GRPO, and inference.

C Details of Temporal Span Alignment

Gaze datasets. We use two eye-tracking datasets built on MIMIC-CXR images, and retain only images that also appear in MIMIC-CXR-VQA. EGD-CXR (Karargyris et al., 2020) contains 1,083 CXRs, covering normal, pneumonia, and congestive heart failure cases in equal proportion, each read by a single radiologist, with a dictation transcript timestamped at the phrase level. REFLACX (Bigolin Lanfredi et al., 2022) contains 3,032 gaze-annotated readings of 2,616 CXRs by five radiologists, with a transcript timestamped at the word level. It was collected in three phases (Table A1). In phases 1 and 2, all five radiologists read a shared set of CXRs, whereas in phase 3, each radiologist read a separate set. In this work, we use only phase 3 of REFLACX for gaze targets, since it provides substantially more distinct images, and use phases 1 and 2 only to examine inter-reader variation (Appendix A).

Non-gaze QA prompt

<image>
{Question}
Your task:

1. Provide your answer inside <answer>...</answer> tags.
2. Do not include any reasoning, explanation, or text outside the <answer> tags.

Figure A2: Prompt for QA pairs without gaze annotation.

Gaze-annotated QA prompt

<image>
You are a radiology assistant reading a chest X-ray.
{Question}
Your task is to output, in this exact order:

1. Identify the anatomical region(s) relevant to the question, inside <organ>...</organ> tags, as name: [x1, y1, x2, y2] separated by semicolons. Output none if no specific anatomical region is relevant.
2. Output the 2 points on this image that determine the answer to this question, inside <gaze>...</gaze> tags, as [x, y], [x, y]. Pick the places that decide the answer, not the centre of the region. Choose two DIFFERENT places.
3. Give your answer inside <answer>...</answer> tags.

All coordinates are integers in $[0, 1000)$, normalised to the image: $(0, 0)$ is the top-left corner, $(999, 999)$ the bottom-right, whatever the image's pixel size. No text outside these tags.

Figure A3: Prompt for gaze-annotated QA pairs.

Matching procedure. To match each QA pair with the radiologist's gaze, we extract the parts of the transcript in which the radiologist discusses what the question asks about. To achieve this, we primarily apply keyword matching between the question and the transcript, and query Qwen3.6-27B only for pairs that keyword matching does not resolve. The LLM receives the question and the timestamped transcript, and returns either a set of time intervals $(t_{\text{start}}, t_{\text{end}})$ or no_match, following the prompt in Figure A4. It has no access to the image or the fixations, and none of its output enters training as text or coordinates. All fixations recorded within the returned intervals are retained, and QA pairs without any matched interval receive no gaze. Table A2 summarizes the resulting coverage.

D Details of Organ Target Construction

Each MIMIC-CXR-VQA question is generated from a Chest ImaGenome scene graph and contains its template arguments (region, finding, or finding category), from which we derive the organ target (Table A3). We map each region to its bounding box using the scene graph metadata. If a region is named, we use its bounding box directly. Regions without a box in Chest ImaGenome are mapped to an adjacent region, preserving laterality (e.g., shoulder $\rightarrow$ clavicle, carina $\rightarrow$ trachea). If only findings or finding categories are named, we take the union of the two most frequent regions per item in a finding-to-region prior. We do not locate findings using the annotations of the given image, since such regions exist for 864 of 910 presence questions answered 'yes' but only 1 of 921 answered 'no', so their existence would reveal the answer. Instead, the prior assigns each finding the regions where it most frequently appears, rather than its bounding box in the given image. Questions without an anatomical referent (e.g., gender, plane) are assigned no region.

Prompt for transcript matching

You are an AI assistant specialized in biomedical topics, especially chest X-ray interpretation and radiology reporting. You are provided two datasets: (1) A transcription spoken by a radiologist, with word-level timestamps. Eye-gaze coordinates are recorded at the same timepoints. (2) A chest X-ray VQA dataset (question–answer pairs). Your goal is to identify transcription spans that correspond to each question–answer pair. The matched time window will later be used to retrieve the radiologist's gaze during that span. Below are requirements for generating the eye-gaze augmented chest xray vqa dataset:

1. Use ONLY the transcription content. Do not use or infer from gaze coordinates, spatial regions, image content, or medical plausibility.
2. Match spans that explicitly or clearly paraphrase the question's target concept, such as a finding, anatomical structure, device, or attribute.
3. Preserve anatomical specificity and laterality. For example, a question about the "right lower lobe" should only match spans referring to the right side, lower lobe, right base, or an equivalent specific region, not generic lung statements.
4. Match by concept relevance, not answer agreement. Negated or uncertain mentions are valid matches. For example, "no consolidation" and "consolidation cannot be excluded" are valid matches for a question about consolidation, even if their polarity differs from the VQA answer.
5. Handling self-corrections: When the radiologist self-corrects (via "actually", "scratch that", "i mean", "sorry", "wait"), decide what to include based on whether the two statements refer to the SAME anatomical region/object:
   (a) SAME region — same region but a different interpretation, refinement, or verbal slip (e.g., "left consolidation → left effusion", "left lung I mean right lung", "macro pacer sorry, cardiac pacer"): include BOTH segments as one continuous span.
   (b) DIFFERENT region — the corrected statement points to a different region, or removes a claim about a specific region that the original included (e.g., "normal lungs → left base opacity", "bilateral effusions → no left effusion"): use ONLY the corrected statement.
6. A VQA pair may match multiple non-contiguous spans. Return all clearly relevant spans.
7. For broad abnormality questions, match all clinically relevant finding statements, including normal findings and explicit absence of findings. Exclude fillers, administrative comments, and purely technical statements unless they affect interpretation.
8. If no span clearly supports the VQA pair, return "no_match". Do not force weak matches.

Figure A4: Prompt template for LLM-based transcript span matching.

Table A1: Statistics of the gaze datasets EGD-CXR and REFLACX.

Dataset	# Readers	# CXRs
EGD-CXR	1	1,083
REFLACX phase 1	5	59
REFLACX phase 2	5	50
REFLACX phase 3	1	2,507

E Details of Gaze Target Construction

Fixation heatmaps are computed on a 48×48 grid over the image. REFLACX records ρ_i per fixation, which varies with the zoom level. EGD-CXR does not record it, so we use a constant $\rho_i = 118$ px/deg. To choose the number of gaze points, we count the modes of each heatmap h as the 8-connected regions of $\{u \in \mathcal{U} : h(u) \geq \frac{1}{2} \max_{u'} h(u')\}$, following the definition of density clusters as connected components of a high-density region (Rinaldo & Wasserman, 2010). The median is two for both h_q and $\tilde{h}_q$.

Table A2: Number of CXRs and QA pairs at each stage of the alignment. *w/ gaze* counts the CXRs and QA pairs, and *Matched* counts those that receive a gaze target after transcript matching. Since gaze is not used at inference, we do not align the test split.

	MIMIC-CXR-VQA		**w/ gaze**		**Matched**	
Split	# CXRs	# QA	# CXRs	# QA	# CXRs	# QA
Train	133,687	290,031	2,199	4,685	1,532	2,494
Valid	8,610	73,567	150	1,328	148	723
Test	500	13,793	9	248	–	–

Table A3: Organ target construction for the 2,494 training QA pairs with gaze.

Question	**Target**	**# QA**
Region named (with box)	Box of the region	1,434
Region named (no box)	Box of an adjacent region	68
Finding/category only	Top-2 prior regions per item	781
No anatomical referent	–	211

F Training Data of Baselines

Table A4 presents the training data of baselines that overlap with our evaluation benchmarks.

G Full Results with Standard Deviation

To assess the stability of fine-tuning, we report the mean and standard deviation over three runs for all fine-tuned VLMs on MIMIC-CXR-VQA (Table A5) and on SLAKE-CXR and VQA-RAD-CXR (Table A6). On MIMIC-CXR-VQA, our method achieves the highest mean on all metrics, with standard deviations comparable to those of Qwen2.5-VL-7B and Gaze-VLM. One of the three LLaVA-Med-7B runs collapsed during fine-tuning, which accounts for its large standard deviation.

Table A4: Training-data overlap between domain-specific baselines and our evaluation benchmarks.

Model	MIMIC-CXR-VQA	SLAKE	VQA-RAD	Other relevant training data
MedGemma-4B		✓	✓	MIMIC-CXR (231K images & reports)
HuatuoGPT-Vision-7B				PubMedVision
LLaVA-Med v1.5				PMC-15M
MedVLM-R1		✓	✓	–
MediX-R1		✓	✓	MIMIC-CXR report generation
CX-Mind	✓			23 datasets (708K images, 2.6M samples)
CheXOne	✓	✓	✓	CheXinstruct-v2, 30 CXR datasets
LLaVA-Rad				MIMIC-CXR (images & reports)
CheXagent	✓		✓	CheXinstruct

Table A5: Mean and standard deviation of fine-tuned VLMs on MIMIC-CXR-VQA over 3 runs. We report accuracy for 'verify' and F1 for 'choose' and 'query' questions. The Acc. and F1 columns denote overall metrics, computed over all three question types and over 'choose' and 'query' questions, respectively. **Bold** and underlined denote the best and second-best mean among fine-tuned VLMs.

Model	**#Param**	**Acc.**	**F1**	**Verify**	**Choose**	**Query**
LLaVA-Med-7B	7B	50.68±28.44	56.05±18.90	53.62±43.66	58.37±19.98	55.39±18.60
Qwen2.5-VL-7B	7B	65.46±0.19	64.45±0.39	77.47±0.36	65.85±0.63	64.05±0.42
Gaze-VLM	7B	65.43±0.14	64.19±0.22	77.47±0.52	65.03±1.23	63.95±0.24
CheXGaze (Ours)	**7B**	**66.18**±0.34	**65.24**±0.53	**78.87**±0.12	**65.92**±0.92	**65.05**±0.64

Table A6: Mean and standard deviation of fine-tuned VLMs on SLAKE-CXR and VQA-RAD-CXR over 3 runs. We report accuracy for closed-ended and recall for open-ended questions. **Bold** and underlined denote the best and second-best mean among fine-tuned VLMs.

		SLAKE-CXR		**VQA-RAD-CXR**	
Model	**#Param**	**Closed**	**Open**	**Closed**	**Open**
LLaVA-Med-7B	7B	84.50±1.34	82.09±0.21	80.17±0.86	61.87±3.37
Qwen2.5-VL-7B	7B	84.21±1.52	82.27±1.08	**85.34**±0.87	57.03±1.26
Gaze-VLM	7B	77.20±1.52	73.88±1.81	81.32±0.50	55.46±0.50
Ours	**7B**	**84.79**±1.01	**85.03**±0.29	**85.34**±1.50	**65.73**±2.47