CARE-X: Towards Clinically Useful Radiology VLMs with Auxiliary Supervision, Reward-Aligned Learning, and Tool-Augmented Measurement

Mercy Prasanna Ranjit MERANJIT@MICROSOFT.COM
Microsoft Research India

Anirban Porya T-ANIPORYA@MICROSOFT.COM
Microsoft Research India

Sathvik Joel T-SATHVIKK@MICROSOFT.COM
Microsoft Research India

Niharika Vadlamudi T-NVADLAMUDI@MICROSOFT.COM
Microsoft Research India

Nikhilesh Chowdary Eathamukkala T-NIKHILESHC@MICROSOFT.COM
Microsoft Research India

Prasanth V V T-PRASANTHVV@MICROSOFT.COM
Microsoft Research India

Abhyuday Kumara Swamy ABHYUDAY.KUMARASWAMY@NARAYANAHEALTH.ORG
Medha AI, Narayana Health, India.

Pranay Narhari Umredkar PRANAYNARHARI.UMREDKAR@NARAYANAHEALTH.ORG
Medha AI, Narayana Health, India.

Pradeep Narayan PRADEEP.NARAYAN.DR@NARAYANAHEALTH.ORG
RTIICS, Narayana Health, India

Vivek Rajagopal VIVEK.RAJAGOPAL@NARAYANAHEALTH.ORG
Medha AI, Narayana Health, India.

Tanuja Ganu TANUJA.GANU@MICROSOFT.COM
Microsoft Research India

Abstract

A clinically useful chest X-ray system must go beyond fluent report generation: it should classify findings with tunable decision thresholds, localize them spatially, and derive the anatomical measurements on which many diagnoses depend. Today's Vision-Language Models (VLMs) treat these as separate problems, if they address them at all—leaving a gap between what radiologists need and what generative models provide. We introduce CARE-X, a chest X-ray VLM that narrows this gap by unifying auxiliary discriminative supervision with reward-aligned generation. CARE-X augments its generative backbone with focal-loss classification and composite-loss grounding heads, co-trained with the language-modeling objective. This auxiliary supervision produces discriminative diagnostic predictions with tunable decision thresholds and precise spatial localization while also improving report quality—evidence that structured prediction and generation reinforce one another. Building on this foundation, Decoupled Clip and Dynamic sAmpling Policy Optimization

(DAPO) leverages task-specific reward signals for report generation, VQA, and spatial grounding, directly optimizing the clinical quality metrics that matter in practice. The result is state-of-the-art performance on the majority of metrics across four report generation benchmarks, 94.0% VQA accuracy on ReXVQA (+6.0 pp over the next-best baseline), and generative spatial decoding that reaches near-parity with dedicated detection heads. Separately, to address measurement-dependent diagnoses, we couple Qwen3-VL-4B-Instruct—an off-the-shelf VLM with native tool-calling capabilities—with deterministic measurement tools while retaining full visual access to the image. This hybrid inference yields +43.6 pp average F1 over perception-only baselines across five measurement-dependent conditions. We validate on rare, high-acuity ICU pathologies using clinical data from Narayana Health (NH), India, and on organ-enlargement conditions with CT-confirmed ground truth, showing that measurement-augmented CXR screening can identify high-risk cases who may require confirmatory imaging.

1. Introduction

Chest X-rays (CXR) are the most commonly performed diagnostic imaging examinations worldwide, serving as the first-line assessment for a wide range of cardiopulmonary conditions across emergency, inpatient, and outpatient settings. Their clinical interpretation is inherently multi-faceted: a radiologist must detect pathologies, localize findings to specific anatomical regions, quantify measurements such as the cardiothoracic ratio, and synthesize all observations into a structured clinical report. This complexity makes chest X-ray interpretation both a compelling benchmark and a demanding real-world target for vision-language models (VLMs). Recent radiology VLMs (Sellergren et al., 2025; Zhou et al., 2026; Zhang et al., 2026) have demonstrated impressive report fluency, yet fluency alone does not constitute clinical fidelity.

1. **No tunable diagnostic thresholds.** Generative VLMs predict diagnoses as free-text tokens without discriminative probability scores that can be thresholded, offering no mechanism to tune sensitivity–specificity trade-offs across clinical contexts. Discriminative models provide these properties but lack the flexibility of open-ended generation.
2. **Cross-entropy (CE) loss does not optimize clinical fidelity.** CE loss optimizes token likelihood but is agnostic to clinical correctness: it penalizes a coordinate error no more than a benign word substitution, treats a "Yes"/"No" inversion identically despite opposite clinical meaning, and weighs the omission of a life-threatening finding the same as an insignificant one.
3. **No capability for measurement-based diagnosis.** Diagnoses such as cardiomegaly, mediastinal widening, and aortic enlargement depend on precise measurements against defined thresholds. No current VLM can perform these measurements; the only prior tool-based approach (Lee et al., 2026) relies on a text-only LLM. A tool-calling VLM enables the interleaving of visual perception and deterministic measurement tools as needed.
4. **No clinical evaluation on rare ICU conditions in non-Western data.** No clinical evaluations have targeted rare, critical ICU conditions—such as pneumoperitoneum, mediastinal shift, fractures, and pneumothorax—using Indian clinical data. Existing VLMs are trained predominantly on Western datasets, and their generalizability to rare conditions in underrepresented populations has not been widely studied.

We address the first three gaps through three complementary strategies, and close the fourth through a dedicated clinical evaluation.

1.1. Contributions

1. **Auxiliary supervision for radiology VLMs.** We present **CARE-X**, a radiology VLM that co-trains focal-loss classification heads (pathology presence, abnormal tube/line placement) and composite-loss grounding heads alongside the autoregressive objective. Auxiliary supervision improves generative performance on the same tasks.
2. **Multi-task reinforcement learning via DAPO.** We apply DAPO (Yu et al., 2025) to CARE-X across report generation, spatial grounding, and closed VQA with task-specific clinical reward signals that directly optimize quality dimensions cross-entropy training cannot capture.
3. **Tool-integrated quantitative reasoning.** We enable tool-calling in a VLM setting with Qwen3-VL-4B-Instruct (Bai et al., 2025) for five measurement-dependent conditions, interleaving visual perception with deterministic measurement tools and benchmarking against perception-only baselines across general-purpose and radiology-specific VLMs.
4. **Validation on Narayana Health (NH) clinical data.**
   We evaluate on rare, high-mortality ICU conditions—pneumoperitoneum, fractures, mediastinal shift, pneumothorax, and abnormal tube/line placement—and on measurement-dependent enlargement conditions with CT-confirmed ground truth, both using NH hospital data, validating CXR-based quantitative screening as a mechanism to flag patients requiring confirmatory imaging.

CARE-X-RL achieves state-of-the-art on the majority of metrics across four report generation benchmarks (MIMIC-CXR, IU-Xray, CheXpert-Plus, ReXGradient), with gains confirmed by CRIMSON (Baharoon et al., 2026), a held-out clinical metric never used as a reward signal. On VQA, it reaches 94.0% accuracy on ReXVQA (+6.0 pp over CheXOne-R1 (Zhang et al., 2026)), while DAPO brings generative spatial decoding to near-parity with the auxiliary detection head. Tool-augmented measurement adds +43.6 pp average F1 over perception-only inference across five conditions.

Generalizable Insights about Machine Learning in the Context of Healthcare

This work yields three insights that extend beyond chest X-ray interpretation. (i) Co-training task-specific auxiliary heads with a generative VLM is mutually reinforcing: the structured supervision enriches shared representations, improving generative outputs on the same tasks—a design principle applicable to any medical domain where both thresholdable discriminative predictions and flexible text generation are needed. (ii) Token-level cross-entropy loss is fundamentally misaligned with clinical quality; reinforcement learning with task-specific rewards can close this gap and bring autoregressive decoding to parity with dedicated structured prediction heads. (iii) For diagnostic tasks defined by quantitative criteria, the decisive factor is not perception versus computation but the division of labour between them: the VLM perceives the radiograph to identify structures and context, while deterministic tools perform the measurement and threshold comparison that neural

networks only approximate – a principle likely to hold across imaging domains where diagnoses hinge on thresholds.

2. Related Work

Recent radiology VLMs, including MedGemma (Sellergren et al., 2025), MedVersa (Zhou et al., 2026), RadVLM (Deperrois et al., 2025), and CheXOne (Zhang et al., 2026), have advanced report generation, VQA, and grounding through multi-task SFT and, in CheXOne's case, GRPO-based RL across all three tasks, yet all remain purely generative and lack discriminative prediction heads with tunable decision thresholds. Rad-Phi4-Vision-CXR (Ranjit et al., 2025) is the closest architectural prior, introducing focal-loss classification and grounding heads alongside a generative backbone, though these heads are trained independently rather than co-trained with the generative objective.

On the RL side, UniRG-CXR (Liu et al., 2026) and RadVLM (Gundersen et al., 2025) apply GRPO to report generation and grounding, but both initialise from purely generative SFT checkpoints without discriminative pre-training. For quantitative reasoning, CXReasonAgent (Lee et al., 2026) couples an LLM with CheXStruct diagnostic tools (Lee et al., 2025), but the backbone never observes the image directly, precluding joint visual–quantitative inference. No prior system unites auxiliary-head co-training, reward-aligned learning, and visual tool use within a single study; a detailed survey is provided in Appendix L.

3. Method

Closing the gap between fluent generation and clinical fidelity requires addressing three distinct failure modes: diagnostic classification in a purely generative setting lacks the deterministic output needed for a tunable decision threshold; spatial localization remains imprecise; and the model cannot perform the quantitative measurements that underpin many radiological diagnoses. We address these through two complementary pipelines. For report generation, classification, and spatial grounding, we train CARE-X—a multimodal architecture with auxiliary task-specific heads—through supervised fine-tuning (SFT) followed by reward-aligned reinforcement learning via DAPO (Section 3.1; Figure 1). For quantitative diagnostic reasoning, we deploy **Qwen3-VL-4B-Instruct** (Bai et al., 2025) as an off-the-shelf VLM with native tool-calling capabilities, augmented with deterministic measurement tools at inference time and requiring no task-specific training (Section 3.2; Figure 2). Table 1 summarizes all tasks, their training paradigms, and optimization objectives.

3.1. CARE-X: Auxiliary Supervision and Reward-Aligned Learning

We first describe the architecture, two-phase training pipeline, loss functions, and inference settings of our trained radiology VLM, which addresses report generation, visual question answering, and spatial grounding.

Table 1: Overview of tasks, training methodology, and optimization objectives. **SFT**: Supervised Fine-Tuning. **DAPO**: Decoupled Clip and Dynamic sAmpling Policy Optimization. **CLM**: Causal Language Modeling loss. Geometric Measurement uses inference-time tool augmentation with no task-specific training.

Task	Approach	Training	SFT Loss	DAPO Reward
Report Generation	Generative	SFT + DAPO	CLM	BERTScore + RadGraph + GREEN
Differential Diagnosis	Generative	SFT + DAPO	CLM	Binary reward (+1 / 0)
Negation Assessment	Generative	SFT + DAPO	CLM	Binary reward (+1 / 0)
Geometric Information Assessment	Generative	SFT + DAPO	CLM	Binary reward (+1 / 0)
Location Assessment	Generative	SFT + DAPO	CLM	Binary reward (+1 / 0)
Abnormality Presence	Generative + Aux. Head	SFT + DAPO	CLM + Focal	Binary reward (+1 / 0)
Abnormality Labels	Generative	SFT	CLM	–
Tubes & Lines Labels	Generative	SFT	CLM	–
Abnormality Location	Generative	SFT	CLM	–
Tubes & Lines Abnormal Placement	Generative + Aux. Head	SFT	CLM + Focal	–
Grounding	Generative + Aux. Head	SFT + DAPO	CLM + GIoU + L1 + Focal	mIoU + GIoU + Box Count
Geometric Measurement	Tool-Augmented	– (inference only)	–	–

† The prompts used for the various tasks are mentioned in Appendix H.

3.1.1. Model Architecture

CARE-X adopts a modular multimodal architecture (Figure 1) consisting of a SigLip2-so400M vision encoder and a Phi-4-mini-instruct (3.8B) autoregressive backbone. The vision encoder is fine-tuned on radiology image–report pairs and interfaced with the language model through a two-layer MLP adapter. Three task-specific auxiliary heads—two binary classifiers for pathology and tubes and line abnormal placement, and a bounding-box detection head for grounding—are attached to the LLM's final hidden layer, providing structured supervision during SFT while leaving auto-regressive text as the primary inference output. Full component-level details are in Appendix A.

3.1.2. Training Pipeline

Supervised Fine-Tuning (SFT). The SFT stage proceeds in three phases: (i) training the image encoder with SigLip2 contrastive loss (Tschannen et al., 2025) on paired chest X-rays and reports, (ii) training the vision–language adapter and auxiliary heads with the encoder and language decoder frozen, and (iii) fine-tuning using LoRA (Hu et al., 2022), with the image encoder frozen and only the adapter, auxiliary heads, and LoRA parameters updated.

Training uses a curated instruction-tuning dataset of ∼5M samples from public chest X-ray datasets, spanning reports, labels, and bounding boxes, where each sample pairs an image with a task-specific user–assistant interaction. More details are provided in Appendix D.

SFT Losses. All tasks share a causal language modeling (CLM) loss on the generative backbone. Tasks with auxiliary heads (Table 1) receive additional supervision: focal loss for the classification heads (pathology presence and tubes and line abnormal placement), addressing the severe class imbalance in binary diagnostic labels; and a weighted composite spatial loss combining Generalized Intersection over Union (GIoU), Mean Intersection over Union (MIoU), L1 coordinate regression, and confidence scoring for the detection head.

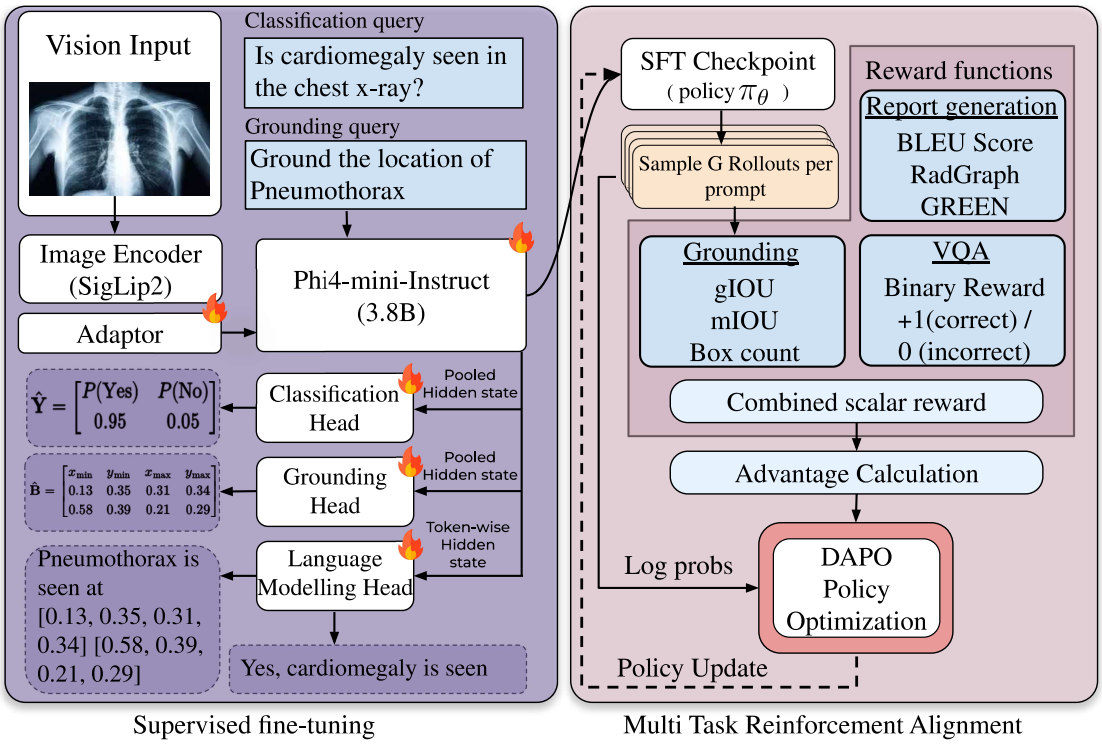


Figure 1: **The CARE-X model. (Left)** Supervised fine-tuning with three task-specific heads for classification, grounding, and report generation/VQA. **(Right)** DAPO with task-specific rewards. See Section 3.1.1 for details.

Decoupled Clip and Dynamic sAmpling Policy Optimization (DAPO). In the second phase, we apply DAPO (Yu et al., 2025) to directly optimize task-specific reward signals that capture clinical quality dimensions. We prefer DAPO over GRPO (Shao et al., 2024) for two properties relevant to our multi-task setting: Clip-Higher (asymmetric clipping with $\varepsilon_{\text{high}} > \varepsilon_{\text{low}}$) prevents the entropy collapse observed in standard GRPO training, and token-level loss normalization ensures equitable gradient contribution across tasks with variable output lengths (short VQA answers vs. long radiology reports).

DAPO Rewards. Because supervised losses optimize token-level likelihood rather than clinical correctness, each DAPO task family uses reward signals that directly target clinical quality:

- **Report Generation:** We use BERTScore (Zhang et al., 2020), RadGraph (Delbrouck et al., 2022), and GREEN (Ostmeier et al., 2024) as the rewards. BERTScore rewards semantically equivalent phrasings that lexical metrics miss; RadGraph rewards F1-level agreement on clinical entities—findings, anatomical locations, and their attributes—between the generated and reference report; GREEN scores each report as the ratio of matched clinical findings to matched findings plus clinically significant errors (false find-

ings, missing findings, wrong location, wrong severity), directly penalizing hallucinations and omissions that would alter clinical decision-making.

- **Closed VQA:** A binary reward assigning +1 for correct and 0 for incorrect responses, directly optimizing diagnostic accuracy where a single-token error inverts the clinical interpretation.
- **Grounding:** A composite spatial reward combining mean Intersection over Union (mIoU) for overlap quality, Generalized IoU (gIoU) to penalize overly large enclosing boxes and a box count reward for correct enumeration of findings. This reward enforces consistency between the number of predicted bounding boxes and the expected count for the target finding or phrase, discouraging both over- and under-detection.

Full training configuration for DAPO along with reward function implementation details can be found in Appendix I.

3.1.3. Inference Settings

The model supports two complementary inference paradigms, corresponding to the approach column in Table 1:

Generative Inference. In generative mode—the primary evaluation paradigm across all tasks—the model produces outputs through auto-regressive decoding: report generation yields free-text clinical findings; closed VQA generates short categorical answers (e.g., `Yes`/`No`); open VQA generates short descriptive answers;and grounding decodes normalized bounding box coordinates $[x_{\min}, y_{\min}, width, height]$ as text tokens.

Auxiliary Head Inference. For tasks with co-trained auxiliary heads—abnormality presence, tubes and lines abnormal placement, and grounding—a second inference path produces structured predictions from the task-specific heads. Classification heads output discriminative probability scores that can be thresholded for binary decisions, while the detection head predicts up to 4 bounding boxes with associated confidence scores. The details regarding the prompts used for the different tasks are mentioned in Appendix H.

3.2. Tool-Augmented Quantitative Measurement

Certain diagnostic tasks—such as determining cardiomegaly from the cardiothoracic ratio or mediastinal widening from the mediastinal-to-thoracic ratio —require precise quantitative measurements that current VLMs cannot reliably derive through perception alone. Tool-augmented measurement operates entirely at inference time: we use Qwen3-VL-4B-Instruct (Bai et al., 2025) with no task-specific fine-tuning, augmenting it with deterministic measurement tools invoked through structured function calling.

3.2.1. Measurement Tools

Each diagnostic condition is associated with a chain of specialized tools that implement clinically grounded measurement criteria from the CheXStruct framework (Lee et al., 2025). The tool chain typically consists of three stages: (i) landmark detection—image-consuming tools that receive the CXR, load CXAS segmentation masks (Seibold et al., 2023), and identify anatomical boundaries; (ii) measurement computation—compute-only tools that calculate

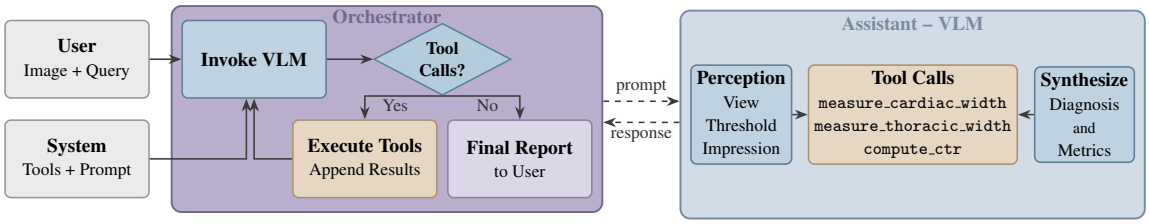


Figure 2: Quantitative Reasoning Inference Pipeline

width ratios or angles from the detected landmarks; and (iii) threshold-based classification—comparison against evidence-based thresholds to produce a normal/abnormal determination.

For example, cardiomegaly assessment follows the chain `measure_cardiac_width` → `measure_thoracic_width` → `compute_ctr`: the first two tools load CXAS segmentation masks (heart, right lung, left lung) to extract normalized anatomical widths, and the final tool computes the cardiothoracic ratio, classifying as abnormal when CTR ≥ 0.50 (PA) or ≥ 0.55 (AP). All measurements use normalized image coordinates, eliminating dependence on pixel spacing metadata (Table 28). Tool chains for the remaining conditions are summarised in Table 24.

3.2.2. Multi-Round Inference

At inference time (Figure 2), a Python orchestrator mediates a multi-turn loop between the VLM—served by vLLM via an OpenAI-compatible API—and a suite of locally executable measurement tools. Given a frontal chest radiograph with condition-specific diagnostic criteria and tool schemas, the VLM reasons over the image and emits structured tool-call requests when quantitative evidence is needed. The orchestrator executes each tool and appends the JSON result to the conversation history; the VLM then reassesses whether further measurements are required. The loop terminates once the model returns a natural-language diagnosis with no pending tool calls, subject to a hard cap of 10 rounds (typically converging within 3).

3.3. Cohort Selection: Narayana Health (NH) Clinical Dataset

Evaluations on public benchmarks such as MIMIC-CXR demonstrate metric-level performance but do not assess whether a model generalizes to the deployment conditions involving rare, high-acuity findings in non-Western patient populations where training data are scarce and missed diagnoses carry severe clinical consequences. To address this gap, we conduct two retrospective evaluation studies on de-identified chest radiographs sourced from NH, a hospital network in India. Both studies are observational and involve no patient intervention; all data were de-identified prior to analysis. This study was conducted under IRB approval (protocol no. NHRTIICSEC/INV/Non-Reg/2026/004).

3.3.1. Study 1: Inpatient and ICU Conditions

The first study assesses CARE-X-SFT's ability to detect rare, high-acuity radiographic findings critical in intensive care and inpatient settings, where delayed or missed diagnosis carries significant clinical risk. The evaluation cohort comprises 1,047 de-identified chest radiographs, each annotated for five binary conditions by qualified radiologists (Table 15,Appendix F). The prevalence of the condition ranges from 2.6% to 5.2% (see Figure 3, reflecting realistic clinical distributions for rare conditions in which the vast majority of radiographs are negative for any given finding. Notably, such rare conditions pose a particular challenge for VLMs that lack sufficient exposure to low-frequency pathological patterns; in this regime, the visual signal is subtle and underrepresented, making it difficult for standard generative models to reliably capture discriminative features. This low-prevalence setting specifically tests the benefit of the dedicated classification head described in Section 3.1.1.Per-condition decision thresholds were selected by sweeping operating points on this cohort to maximise balanced sensitivity–specificity, and the same cohort is used for the reported metrics; we quantify the resulting selection optimism by 3-fold cross-validation in Appendix F.1.

3.3.2. Study 2: Outpatient Aorta Enlargement Conditions

The second study evaluates whether tool-augmented VLM inference—in which the model invokes external measurement tools while retaining full visual access to the radiograph—can identify high-risk patients with organ or vascular enlargement who may require confirmatory cross-sectional imaging. The evaluation cohort consists of 122 positive cases across three measurement-dependent conditions (Table 16). Critically, ground truth for all positive cases has been confirmed by CT imaging. This CT-confirmed design distinguishes our evaluation from prior CXR studies that rely solely on radiologist consensus, which can be subjective for borderline enlargement findings, and enables a rigorous assessment of whether measurement-based CXR screening can serve as an effective triage step to identify patients warranting confirmatory imaging.

4. Experimental Design

We evaluate the three strategies described in Section 3 through controlled comparisons. All learning-based experiments build on CARE-X; tool-augmented measurement experiments use Qwen3-VL-4B-Instruct without task-specific training.

4.1. Auxiliary Supervision and DAPO Evaluation

To isolate the contribution of auxiliary supervision, we compare three model variants: (i) a baseline generative model trained without auxiliary heads, (ii) a co-trained model evaluated in generative mode, where predictions are inferred from generated text, and (iii) the same co-trained model evaluated via its auxiliary head outputs. This setup disentangles whether improvements arise from better internal representations benefiting generation, or from the explicit predictive capacity of the auxiliary heads. Building on the best co-trained variant, we further fine-tune with multi-task DAPO (Section 3.1.2), enabling comparison of

SFT-only and SFT+DAPO models across all tasks. Full SFT and DAPO training hyperparameters are reported in Appendix C and Appendix I.1, respectively.

4.2. Tool-Augmented Measurement Evaluation

To evaluate quantitative diagnostic reasoning, we assess the tool-augmented measurement pipeline (Section 3.2) on five conditions summarized in Table 24, with all dataset details in Appendix K.1 & Appendix K.2. We compare three experimental paradigms: (i) tool-augmented measurement via structured function calling; (ii) perception-only inference, where the VLM classifies directly from the image without tool access, evaluated in both a single-image setting and a dual-image setting with an anatomy overlay containing condition-relevant CXAS segmentation masks, as described in Appendix K.1. The dual-image setting tests whether explicit structural cues can reduce the VLM's perceptual ambiguity in perception-only inference; and (iii) external perception baselines (CheXOne and MedGemma) under identical prompts. All configurations are assessed via sensitivity, specificity, and F1.

4.2.1. View-Aware Thresholds

For conditions where the radiographic projection (AP vs. PA) influences anatomical measurements—cardiomegaly and mediastinal widening—we apply view-dependent classification thresholds. We ablate two settings in Tool Augmented Measurements for view classification: (i) model perception, where the VLM infers the radiographic projection directly from the image, and (ii) oracle ground-truth injection, where the true view label is provided to the model, enabling systematic evaluation of how view classification accuracy affects downstream diagnostic performance.

5. Results

We evaluate the proposed framework across three complementary strategies: reward-aligned learning via DAPO (Section 5.1), auxiliary supervision (Section 5.2), and tool-augmented measurement for quantitative diagnostic reasoning (Section 5.3). We further validate clinical generalizability on NH hospital data across rare ICU conditions and measurement-dependent enlargement diagnoses (Section 5.4).

5.1. Reward-Aligned Learning via DAPO

Across report generation, closed VQA, and spatial grounding, two results hold consistently. First, CARE-X-RL achieves state-of-the-art performance against MedGemma (Sellergren et al., 2025), MedVersa (Zhou et al., 2026), and CheXOne-R1 (Zhang et al., 2026), leading on 26 of 32 metric×benchmark combinations across four report generation benchmarks and all five VQA categories against external baselines (Tables 2, 3, 4). Second, CARE-X-RL consistently improves over CARE-X-SFT, isolating the contribution of reward-aligned learning from architectural and data choices. Notably, CARE-X-SFT itself already matches or surpasses the strongest baselines on most tasks, confirming that auxiliary supervision provides a strong foundation; DAPO then pushes performance further, with the largest gains on clinically grounded and composite metrics (1/RadCliQ).

Table 2: Report generation performance across four benchmarks (Findings). Best result per metric is **bolded**; second best is underlined. All metrics are higher-is-better (↑).

Model	BLEU ↑	BERTScore ↑	SEmb ↑	RadGraph ↑	1/RadCliQ ↑	RaTEScore ↑	GREEN ↑	CRIMSON ↑
MIMIC-CXR								
MedGemma	0.165	0.346	0.339	0.159	0.744	0.549	0.293	0.082
MedVersa	0.209	0.448	0.466	0.273	1.103	0.550	0.374	0.170
CheXOne-R1	0.218	0.461	0.455	0.235	1.060	0.519	0.314	0.080
CARE-X-SFT	0.234	0.444	0.439	0.251	1.033	0.584	0.351	–
CARE-X-RL	**0.262**	**0.477**	**0.468**	**0.283**	**1.183**	**0.607**	**0.388**	**0.236**
IU-Xray								
MedGemma	0.217	0.475	0.600	0.260	1.340	0.678	**0.724**	0.524
MedVersa	0.206	0.527	0.606	0.235	1.460	0.650	0.631	0.623
CheXOne-R1	0.265	**0.542**	0.611	0.280	1.669	0.617	0.585	0.490
CARE-X-SFT	0.214	0.430	0.521	0.230	1.068	0.589	0.548	–
CARE-X-RL	**0.272**	0.537	**0.653**	**0.302**	**1.859**	**0.702**	0.630	**0.666**
CheXpert-Plus								
MedGemma	0.147	0.328	0.325	0.137	0.706	0.511	0.246	0.111
MedVersa	0.129	0.323	0.344	0.147	0.719	0.470	0.243	0.086
CheXOne-R1	0.180	**0.430**	**0.487**	0.243	**1.048**	0.522	0.250	0.094
CARE-X-SFT	0.163	0.348	0.448	0.205	0.850	0.541	0.282	–
CARE-X-RL	**0.205**	0.397	0.418	**0.247**	0.934	**0.579**	**0.306**	**0.266**
ReXGradient								
MedGemma	0.200	0.427	0.479	0.223	1.008	0.617	**0.566**	0.282
MedVersa	0.210	0.431	0.498	0.202	1.008	0.527	0.532	0.382
CheXOne-R1	0.229	0.483	0.498	0.210	1.116	0.535	0.428	0.127
CARE-X-SFT	0.295	0.477	0.536	0.311	1.348	0.618	0.492	–
CARE-X-RL	**0.300**	**0.512**	**0.563**	**0.323**	**1.556**	**0.653**	0.506	**0.448**

5.1.1. Report Generation

Setup. We evaluate report generation on four benchmarks: MIMIC-CXR, IU-Xray, CheXpert-Plus, and the private ReXGradient test set. We compare against MedGemma, MedVersa, and CheXOne using baseline results from the ReXrank leaderboard (Zhang et al., 2025b) under the Findings evaluation protocol.

CRIMSON, a held-out metric never used as a reward signal, independently confirms that DAPO improvements reflect genuine clinical quality. CRIMSON (Baharoon et al., 2026) scores only abnormal findings—excluding normals to prevent style-inflation—and penalises errors (hallucinations, omissions, attribute mistakes) in proportion to clinical urgency, yielding a severity-weighted score in $(-1, 1]$. CRIMSON scores are reported for CARE-X-RL only; CARE-X-SFT was not evaluated on this metric as CRIMSON became available after the SFT checkpoint was submitted to ReXrank (Zhang et al., 2025b). CARE-X-RL leads on all four benchmarks, with the largest margins on ReXGradient and CheXpert-Plus. Because CRIMSON is entirely independent of our training rewards, these gains provide strong evidence of transferable quality improvements rather than reward-specific overfitting.

Table 3: Visual question answering accuracy on ReXVQA (↑). Best results are **bolded**.

Model	Negation ↑	Presence ↑	Location ↑	Diff. Diag. ↑	Geometric ↑	Overall ↑
MedGemma	0.898	0.794	0.750	0.819	0.687	0.834
CheXOne-R1	0.911	0.880	0.804	0.845	0.737	0.880
CARE-X-SFT	0.949	0.889	0.819	0.898	**0.760**	0.908
CARE-X-RL	**0.965**	**0.931**	**0.883**	**0.933**	0.749	**0.940**

Table 4: Grounding performance across four benchmarks. All metrics are higher-is-better (↑). Best results are **bolded**

Model	Inference Setting	Anatomy (Chest ImaGenome)		Abnormality (VinDR)		Phrase (Padchest)		Phrase (MS)	
		mAP ↑	mIoU ↑	mAP ↑	mIoU ↑	mAP ↑	mIoU ↑	mAP ↑	mIoU ↑
RadVLM	Generative	0.853	–	**0.495**	**0.358**	0.443	0.288	0.829	0.531
CARE-X-SFT	Generative	0.808	0.535	0.323	0.233	0.613	0.398	0.762	0.494
CARE-X-SFT	Auxiliary head	0.865	0.580	0.404	0.293	**0.676**	**0.451**	0.803	**0.535**
CARE-X-RL	Generative	**0.868**	**0.603**	0.393	0.274	0.660	0.432	**0.817**	0.510

5.1.2. Closed Visual Question Answering

Setup. We evaluate closed-form VQA on the ReXVQA benchmark, comprising 41,007 question–answer pairs across five clinically relevant categories. We compare against MedGemma and CheXOne-R1 (Zhang et al., 2026), the two strongest publicly benchmarked models on ReXVQA; MedVersa does not report results on this benchmark.

The largest gains concentrate in clinically high-risk categories. CARE-X-RL reaches 94.0% overall accuracy (Table 3), with the largest improvements over CheXOne-R1 in differential diagnosis (+8.8 pp), location assessment (+7.9 pp), and negation (+5.4 pp). These categories carry the highest clinical stakes—a negation inversion ("no pneumothorax" when one is present) or a missed diagnosis can misdirect treatment—and are precisely where binary reward outperforms token-level CE loss, which cannot distinguish a one-token error that flips the diagnosis from one that merely changes wording."

5.1.3. Spatial Grounding

Setup. We evaluate spatial grounding across four benchmarks spanning complementary grounding tasks: anatomy localization on Chest ImaGenome, abnormality localization on VinDR-CXR, and phrase grounding on PadChest and MS-CXR. Performance is measured using mean Average Precision (mAP) and mean Intersection-over-Union (mIoU). We compare against RadVLM, a strong grounding-focused baseline, and assess both generative decoding and auxiliary detection head outputs for our models.

DAPO-trained generative output approaches or exceeds the SFT auxiliary detection head. Across four grounding benchmarks (Table 4), DAPO yields consistent mAP improvements of 5–8% over SFT in generative mode, with the largest gain on Chest ImaGenome mIoU (+12.7%). More strikingly, DAPO narrows and in some cases eliminates the gap between generative decoding and the auxiliary detection head: on Anatomy,

Table 5: Closed VQA performance on abnormality classification. (`Th`) denotes the classification threshold. All metrics are higher-is-better (↑). Best results are bolded.

Model	Inference Setting	Sensitivity ↑	Specificity ↑	PPV ↑	NPV ↑	F1 ↑	AUC ↑
CARE-X-SFT	Generative	0.932	0.693	0.895	0.784	**0.913**	-
CARE-X-SFT(Th=0.5)	Auxiliary Head	**0.943**	0.656	0.885	**0.805**	**0.913**	0.868
CARE-X-SFT(Th=0.6)	Auxiliary Head	0.855	**0.813**	**0.927**	0.668	0.89	0.868
CheXOne	Generative	0.878	0.580	0.854	0.629	0.866	-
MedGemma	Generative	0.798	0.712	0.886	0.557	0.839	-

CARE-X-RL generative (0.868 mAP) surpasses the SFT detection head (0.865), and on Phrase grounding (Padchest), the gap shrinks from 0.063 to 0.016 mAP. This result is practically significant: it demonstrates that reward-aligned learning can bring autoregressive spatial decoding to parity with structured prediction, offering clinicians a single generative inference mode without requiring auxiliary heads at test time.

Three task-specific exceptions to this overall picture (competitor leads on certain datasets, a geometric VQA regression under binary reward, and VinDR abnormality grounding trailing a grounding-specialist baseline) are analysed in Appendix J.

5.2. Auxiliary Supervision

Setup. We incorporate lightweight auxiliary heads—two classification heads for abnormality presence and tubes and lines abnormal placement and a detection head for grounding—attached to the language model's hidden states.

Effect of auxiliary grounding supervision. Table 4 shows that the auxiliary detection head consistently outperforms generative decoding. On Chest ImaGenome, mAP and mIoU increase by +5.7 pp and +4.5 pp, while the largest gains occur on VinDR (+8.1 pp mAP, +6.0 pp mIoU). This indicates that the composite spatial loss enhances geometric precision in shared representations, benefiting autoregressive decoding. Appendix B isolates this effect directly: against an otherwise identical generative-only model trained without auxiliary heads, co-training improves generative performance on all eight grounding metrics and on classification F1, confirming that the auxiliary supervision enriches the shared representation rather than merely adding a separate prediction pathway.

Closed VQA: Chest ImaGenome Abnormality Classification. On Chest ImaGenome (Table 5), the generative model achieves high sensitivity (0.932). Adding a classification head improves NPV (0.784 → 0.805 at threshold 0.5), indicating better separation of normal and abnormal cases. The head also enables tunable operating points: at threshold 0.6, specificity (0.813) and PPV (0.927) increase, allowing a trade-off between high-sensitivity screening and high-specificity confirmation within a single model.

Compared to CheXOne and MedGemma, our model achieves higher sensitivity (+5.4%, +13.4%) and NPV (+15.5%, +22.7%), while the classification head at threshold 0.6 achieves the highest specificity and PPV, demonstrating well-calibrated predictions across operating regimes.

Table 6: Perception-only versus tool-augmented measurement (%). Italic Δ = absolute improvement in percentage points.

	Sensitivity			Specificity			F1		
Condition	Perc.	Tool	Δ	Perc.	Tool	Δ	Perc.	Tool	Δ
CM	86.96	96.07	*+9.1*	20.60	93.02	*+72.4*	74.56	96.00	*+21.4*
MW	88.18	95.42	*+7.2*	18.00	99.52	*+81.5*	72.63	97.47	*+24.8*
AK	61.71	99.51	*+37.8*	41.72	100.00	*+58.3*	60.31	99.76	*+39.5*
AAE	27.56	100.00	*+72.4*	82.80	100.00	*+17.2*	39.33	100.00	*+60.7*
DA[†]	20.00	100.00	*+80.0*	50.00	100.00	*+50.0*	28.57	100.00	*+71.4*
Avg Δ			*+41.3*			*+55.9*			*+43.6*

[†]Limited public dataset availability for descending aorta enlargement resulted in only 7 test samples; results should be interpreted with caution.

Closed VQA: Tubes & Lines Abnormal Placement For the closed VQA task of understanding the tubes and lines abnormal placement, we have used RANZCR Seah et al. (2020) dataset for benchmarking. The results regarding this specific task and the inference pipeline is discussed in Appendix E.

5.3. Measurement based Quantitative Reasoning

Complementing the training-time strategies evaluated above, this section examines inference-time tool augmentation as a means to address the quantitative reasoning demands that auxiliary supervision and DAPO do not directly target.

Tool-augmented measurement dramatically outperforms perception-only inference across all five conditions. Table 6 compares tool-augmented inference against perception-only inference with anatomy overlay (Appendix K.1) per condition. The average F1 improvement is 43.6pp, ranging from +21.4pp (CM) to +71.4pp (DA;n=7 see table footnote), with AK and AAE exceeding 96% F1. Residual errors trace to LLM-intrinsic numerical reasoning limitations rather than the tools themselves (see Appendix K.3 & Appendix K.4 respectively).

Effect of anatomy-overlay input on perception. Providing the anatomy overlay of relevant anatomies consistently improves perception-only sensitivity, often dramatically (Table 27). On CM and MW, sensitivity rises by +20.8 and +61.6pp respectively, though with reduced specificity. For AAE and DA, single-image perception yields 0% sensitivity, whereas the overlay activates non-trivial detection—suggesting it provides spatial grounding that prevents the model from defaulting to conservative "normal" predictions.

View misclassification is the primary source of diagnostic error for view-dependent conditions; accurate view information substantially closes the gap to ceiling performance. Table 26 presents the view-classification ablation for CM and MW. For CM, providing the ground-truth view label raises F1 and specificity substantially—access to the correct view enables the model to select AP-appropriate thresholds that reduce false

Table 7: Inpatient and ICU Pathology Classification Under Out-of-Distribution Performance Evaluation. Bold = best per column.

		Fracture		Med. Shift		Pneumop.		Pneumotx.		Tubes & Lines Abn. Placement	
Model	Inference Strategy	Sens	Spec	Sens	Spec	Sens	Spec	Sens	Spec	Sens	Spec
CheXOne	Generative	0.41	0.90	0.80	0.78	0.67	**0.98**	**0.85**	0.72	0.03	**0.97**
MedGemma	Generative	0.05	**1.00**	**1.00**	0.53	0.00	1.00	0.52	0.73	0.18	0.87
CARE-X-SFT†	Auxiliary Head	**0.62**	0.64	0.83	**0.86**	**0.89**	0.94	0.83	**0.75**	**0.66**	0.77

†Decision thresholds tuned per condition: Fracture (0.50), Med. Shift (0.55), Pneumoperitoneum (0.65), Pneumothorax (0.65), Tubes & Lines Abn. Placement (0.40). See Appendix F.1 for a cross-validated threshold-selection analysis.

positives from magnified cardiac silhouettes. For MW, the improvement is even more pronounced: the GT-view oracle achieves the highest F1.

5.4. NH Dataset Evaluation

Study 1 – Inpatient and ICU Conditions: Across all five pathologies, CARE-X-SFT achieves more **balanced and robust performance** than CheXOne and MedGemma (Table 7) under a strict **out-of-distribution (OOD)** evaluation setting. The baselines exhibit characteristic OOD failure modes: CheXOne collapses on Tubes & Lines Abnormal Placement (refer to Appendix E) despite reasonable performance on Pneumothorax and Fracture, while MedGemma fails entirely on Pneumoperitoneum (sensitivity = 0.00) and shows near-zero sensitivity on Fracture, rendering it unreliable for critical findings.

CARE-X-SFT achieves the highest sensitivity in three out of five conditions while maintaining reasonable specificity. The improvements are most pronounced for Pneumoperitoneum and Tubes & Lines Abnormal Placement, where it substantially outperforms both baselines. For Mediastinal Shift, CARE-X-SFT attains strong sensitivity alongside high specificity, in contrast to MedGemma's high-sensitivity bias. Taken together, these results suggest that CARE-X-SFT generalizes more reliably to OOD clinical data, particularly for rare conditions.

Study 2 – Outpatient Measurement-Based Enlargement Conditions: The evaluation cohort comprises exclusively confirmed aortic-enlargement-related positive cases (Table 16, Appendix F), with ground truth established via corresponding CT reports. sensitivity (recall) serves as the primary metric of interest, since the absence of true negatives precludes meaningful estimation of specificity. We evaluate Qwen3-VL-4B-Instruct across perception and measurement (tool-calling) modes (Section 4.2). The measurement-based setting yields a **+10.65%** improvement in disease classification recall over the perception-only baseline (Table 8), corroborating our central findings (Table 27).

6. Discussion

Our results establish that discriminative and generative objectives are mutually reinforcing. Co-training auxiliary classification and grounding heads with the language-modeling objective improves generative report quality and spatial localization, and DAPO amplifies this further—bringing generative spatial decoding to near-parity with the SFT detection head, which suggests a single generative inference mode can subsume dedicated structured prediction at deployment.

Table 8: Perception vs. Tool-Augmented Performance on Enlargement Conditions (NH Outpatient Cohort).

Model	Overlay	Recall
Perception-Only		79.51%
Perception-Only	✓	83.61%
Tool-Augmented		**94.26**%

The classification head offers a distinct deployment advantage: discriminative probability scores with tunable thresholds enable clinicians to shift between high-sensitivity screening and high-specificity confirmation from a single forward pass. On the NH ICU evaluation, CARE-X-SFT is the only model that achieves consistently meaningful likelihood ratios across all five rare conditions, with bidirectional discriminability (e.g., LR+ of 14.83 and LR- of 0.12 for pneumoperitoneum) enabling actionable revision of post-test probability. The clinical significance lies in uniform reliability across the diagnostic spectrum of an adult cardiac ITU, maintained under distribution shift (Appendix G).

The second key insight concerns the boundary between what should be learned and what should be computed. Measurement-dependent diagnoses—cardiomegaly, mediastinal widening, aortic enlargement—are fundamentally different from pattern-recognition tasks: they reduce to whether a ratio exceeds a clinically defined threshold. Our results show that no perception-only VLM, regardless of size or medical specialization, can reliably make these determinations through visual approximation alone, whereas coupling a general-purpose VLM with deterministic measurement tools yields dramatic improvements (averaging +43.6 pp F1 across five conditions). Crucially, unlike prior tool-augmented approaches that rely on text-only LLMs (Lee et al., 2026), our pipeline retains full visual access to the radiograph, enabling the VLM to reason about what requires measurement while delegating the how to specialized tools. The clinical value of this approach is validated on CT-confirmed enlargement cases from NH, where tool-augmented CXR screening demonstrates the ability to identify high-risk cases who may require confirmatory imaging.

7. Limitations and Future Work

Tool-augmented measurements depend on CXAS segmentation masks, whose quality we did not independently validate on the NH cohort—segmentation failures propagate directly into diagnostic error. Tool calling currently uses Qwen3-VL-4B-Instruct rather than CARE-X; unifying it within a single radiology-specialized VLM is an important next step. The DAPO rewards are imperfect proxies for clinical quality: binary VQA reward cannot capture partial correctness in geometric reasoning, and graduated, severity-weighted rewards remain open. All evaluations here are retrospective. Prospective, clinician-in-the-loop validation is a future work.

References

Ayat Abedalla, Malak Abdullah, Mahmoud Al-Ayyoub, and Elhadj Benkhelifa. 2ST-UNet: 2-stage training model using U-Net for pneumothorax segmentation in chest x-rays. In 2020 International Joint Conference on Neural Networks (IJCNN), pages 1–6. IEEE, 2020. doi: 10.1109/IJCNN48605.2020.9207268.

Mohammed Baharoon, Thibault Heintz, Siavash Raissi, Mahmoud Alabbad, Mona Alhammad, Hassan AlOmaish, Sung Eun Kim, Oishi Banerjee, and Pranav Rajpurkar. Crimson: A clinically-grounded llm-based metric for generative radiology report evaluation, 2026. URL https://arxiv.org/abs/2603.06183.

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, 2025. URL https://arxiv.org/abs/2511.21631.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. PadChest: A large chest x-ray image dataset with multi-label annotated reports. Medical image analysis, 66:101797, 2020. doi: 10.1016/j.media.2020.101797.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P. Langlotz. CheXpert Plus: Augmenting a large chest X-ray dataset with text radiology reports, patient demographics and additional image formats, 2024. URL https://arxiv.org/abs/2405.19538.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. Improving the factual correctness of radiology report generation with semantic rewards. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, Findings of the Association for Computational Linguistics: EMNLP 2022, pages 4348–4360, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.319. URL https://aclanthology.org/2022.findings-emnlp.319/.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association, 23(2):304–310, 2016.

Nicolas Deperrois, Hidetoshi Matsuo, Samuel Ruipérez-Campillo, Moritz Vandenhirtz, Sonia Laguna, Alain Ryser, Koji Fujimoto, Mizuho Nishio, et al. RadVLM: A multitask conversational vision-language model for radiology, 2025. URL https://arxiv.org/abs/2502.03333.

Sijing Feng, Damian Azzollini, Ji Soo Kim, Cheng-Kai Jin, Simon P Gordon, Jason Yeoh, Eve Kim, Mina Han, Andrew Lee, Aakash Patel, et al. Curation of the CANDID-PTX dataset with free-text reports. Radiology: Artificial Intelligence, 3(6):e210136, 2021. doi: 10.1148/ryai.2021210136.

Benjamin Gundersen, Nicolas Deperrois, Samuel Ruiperez-Campillo, Thomas M. Sutter, Julia E. Vogt, Michael Moor, Farhad Nooralahzadeh, and Michael Krauthammer. Enhancing radiology report generation and visual grounding using reinforcement learning, 2025. URL https://arxiv.org/abs/2512.10691.

Gregory Holste, Song Wang, Ajay Jaiswal, Yuzhe Yang, Mingquan Lin, Yifan Peng, and Atlas Wang. CXR-LT: Multi-Label Long-Tailed Classification on Chest X-Rays. PhysioNet, 2023. doi: 10.13026/721s-vs37. URL https://doi.org/10.13026/721s-vs37. Version 1.0.0.

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. In International Conference on Learning Representations, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI conference on artificial intelligence, volume 33, pages 590–597, 2019. doi: 10.1609/aaai.v33i01.3301590.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042, 2019. URL https://arxiv.org/abs/1901.07042.

Hyungyung Lee, Geon Choi, Jung-Oh Lee, Hangyul Yoon, Hyuk Gi Hong, and Edward Choi. CXReasonBench: A benchmark for evaluating structured diagnostic reasoning in chest x-rays. In Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track, 2025. URL https://arxiv.org/abs/2505.18087.

Hyungyung Lee, Hangyul Yoon, and Edward Choi. CXReasonAgent: Evidence-grounded diagnostic reasoning agent for chest X-rays, 2026. URL https://arxiv.org/abs/2602.23276.

Qianchu Liu, Sheng Zhang, Guanghui Qin, Yu Gu, Ying Jin, Sam Preston, Yanbo Xu, Sid Kiblawi, Wen-wai Yim, Timothy Ossowski, Tristan Naumann, Mu Wei, and Hoifung Poon. Scaling medical imaging report generation with multimodal reinforcement learning, 2026. URL https://arxiv.org/abs/2601.17151.

Ha Q. Nguyen, Khanh Lam, Linh T. Le, Hieu H. Pham, Dat Q. Tran, Dung B. Nguyen, Dung D. Le, Chi M. Pham, Hang T. T. Tong, Diep H. Dinh, Cuong D. Do, Luu T. Doan, Cuong N. Nguyen, Binh T. Nguyen, Que V. Nguyen, Au D. Hoang, Hien N. Phan, Anh T. Nguyen, Phuong H. Ho, Dat T. Ngo, Nghia T. Nguyen, Nhan T. Nguyen, Minh Dao, and Van Vu. VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. Scientific Data, 9(1):429, 2022. doi: 10.1038/s41597-022-01498-w.

OpenAI. GPT-5 system card, 2025. URL https://openai.com/index/gpt-5-system-card/. Accessed: 2026-07-31.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, and Jean-Benoit Delbrouck. GREEN: Generative radiology report evaluation and error notation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, Findings of the Association for Computational Linguistics: EMNLP 2024, pages 374–390, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.21. URL https://aclanthology.org/2024.findings-emnlp.21/.

Ankit Pal, Jung-Oh Lee, Xiaoman Zhang, Malaikannan Sankarasubbu, Seunghyeon Roh, Won Jung Kim, Meesun Lee, and Pranav Rajpurkar. ReXVQA: A large-scale visual question answering benchmark for generalist chest X-ray understanding, 2025. URL https://arxiv.org/abs/2506.04353.

Mercy Prasanna Ranjit, Anirban Porya, Shaury Srivastav, Niharika Vadlamudi, Nikhilesh Chowdary Eathamukkala, Shashank Udyavar, Rahul Kumar, and Tanuja Ganu. Rad-Phi4-Vision-CXR: A compact multimodal assistant for versatile radiology workflows. In Machine Learning for Health (ML4H), 2025. URL https://www.microsoft.com/en-us/research/wp-content/uploads/2025/11/Rad-Phi4-Vision-CXR-ML4H2025_1.pdf.

Jarrel Seah, Jen, Maggie, Meng Law, Phil Culliton, and Sarah Dowd. RANZCR CLiP - catheter and line position challenge. https://kaggle.com/competitions/ranzcr-clip-catheter-line-classification, 2020. Kaggle.

Constantin Seibold, Alexander Jaus, Matthias A. Fink, Moon Kim, Simon Reiß, Ken Herrmann, Jens Kleesiek, and Rainer Stiefelhagen. Accurate fine-grained segmentation of human anatomy in radiographs via volumetric pseudo-labeling. arXiv preprint arXiv:2306.03934, 2023. URL https://arxiv.org/abs/2306.03934.

Constantin Marc Seibold, Simon Reiß, M. Saquib Sarfraz, Matthias A. Fink, Victoria Mayer, Jan Sellner, Moon Sung Kim, Klaus H. Maier-Hein, Jens Kleesiek, and Rainer Stiefelhagen. Detailed annotations of chest X-rays via CT projection for report understanding. In 33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022. BMVA Press, 2022. URL https://arxiv.org/abs/2210.03416.

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. URL https://arxiv.org/abs/2507.05201.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa,

Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. URL https://arxiv.org/abs/2502.14786.

Maria De La Iglesia Vayá, Jose Manuel Saborit, Joaquim Angel Montell, Antonio Pertusa, Aurelia Bustos, Miguel Cazorla, Joaquin Galant, Xavier Barber, Domingo Orozco-Beltrán, Francisco García-García, et al. BIMCV COVID-19+: a large annotated dataset of RX and CT images from COVID-19 patients. arXiv preprint arXiv:2006.01174, 2020.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2097–2106, 2017. doi: 10.1109/CVPR.2017.369.

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. In Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/hash/17e62166fc8586dfa4d1bc0e1742c08b-Abstract-round2.html.

Justin Xu, Zhihong Chen, Andrew Johnston, Louis Blankemeier, Maya Varma, Jason Hom, William J. Collins, Ankit Modi, Robert Lloyd, Benjamin Hopkins, Curtis Langlotz, and Jean-Benoit Delbrouck. Overview of the first shared task on clinical text generation: RRG24 and "discharge me!". In Dina Demner-Fushman, Sophia Ananiadou, Makoto Miwa, Kirk Roberts, and Junichi Tsujii, editors, Proceedings of the 23rd Workshop on Biomedical Natural Language Processing, pages 85–98, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.bionlp-1.7. URL https://aclanthology.org/2024.bionlp-1.7/.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale, 2025. URL https://arxiv.org/abs/2503.14476.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In International Conference on Learning Representations, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Xiaoman Zhang, Julián N. Acosta, Josh Miller, Ouwen Huang, and Pranav Rajpurkar. ReXGradient-160K: A large-scale publicly available dataset of chest radiographs with free-text reports, 2025a. URL https://arxiv.org/abs/2505.00228.

Xiaoman Zhang, Hong-Yu Zhou, Xiaoli Yang, Oishi Banerjee, Julián N. Acosta, Josh Miller, Ouwen Huang, and Pranav Rajpurkar. ReXrank: A public leaderboard for AI-powered radiology report generation. In Junde Wu, Jiayuan Zhu, Min Xu, and Yueming Jin, editors, Proceedings of The First AAAI Bridge Program on AI for Medicine and Healthcare, volume 281 of Proceedings of Machine Learning Research, pages 90–99. PMLR, 25 Feb 2025b. URL https://proceedings.mlr.press/v281/zhang25b.html.

Yabin Zhang, Chong Wang, Yunhe Gao, Jiaming Liu, Maya Varma, Justin Xu, Sophie Ostmeier, Jin Long, Sergios Gatidis, Seena Dehkharghani, Arne Michalson, Eun Kyoung Hong, Christian Bluethgen, Haiwei Henry Guo, Alexander Victor Ortiz, Stephan Altmayer, Sandhya Bodapati, Joseph David Janizek, Ken Chang, Jean-Benoit Delbrouck, Akshay S. Chaudhari, and Curtis P. Langlotz. A reasoning-enabled vision-language foundation model for chest x-ray interpretation, 2026. URL https://arxiv.org/abs/2604.00493.

Hong-Yu Zhou, Julián Nicolás Acosta, Subathra Adithan, Suvrankar Datta, Eric J. Topol, and Pranav Rajpurkar. MedVersa: A generalist foundation model for diverse medical imaging tasks. NEJM AI, 3(4):AIoa2500595, 2026. doi: 10.1056/AIoa2500595. URL https://ai.nejm.org/doi/full/10.1056/AIoa2500595.

Appendix A. Model Architecture Details

CARE-X follows a modular multimodal architecture comprising four components: (i) a vision encoder, (ii) a vision language adapter, (iii) a small language model (SLM), and (iv) task-specific auxiliary heads.

(i) Vision Encoder. We employ the pre-trained `SigLIP2-so400M-patch14-224` as the image encoder, a 400 million parameter model, which is further fine-tuned on CXR datasets to enhance radiology-specific understanding.

(ii) Vision Language Adapter. A two-layer MLP with GELU activation projects the vision encoder's output embeddings into the hidden dimension of the language model. Weights are initialized with Xavier uniform (gain = 0.1) and zero biases, ensuring stable early training.

(iii) Language Model. We use `Phi-4-mini-instruct` as the autoregressive language backbone, a 3.8 billion parameter model.

(iv) Auxiliary Heads. To provide structured supervision beyond the generative CLM loss, we attach lightweight heads to the LLM's last hidden layer:

- **Classification Head:** Two binary classifiers for abnormality presence detection and tubes/lines abnormal placement detection. Each applies a two-layer projector (Linear $\rightarrow$ ReLU $\rightarrow$ Linear $\rightarrow$ Dropout) followed by a three-layer classification network. Input features are first masked using a prompt-specific mask, then mean-pooled over the prompt tokens to obtain a single representation per sample, which is passed through the classification head to produce a scalar logit. The model is trained using focal loss.
- **Detection Head:** A bounding box regression module for visual grounding. It shares the same projector architecture, followed by separate box regression and confidence heads. The box head predicts up to 4 bounding boxes via sigmoid-activated coordinates $[x, y, w, h] \in [0, 1]^4$, while the confidence head predicts an objectness score per box.

Both heads operate on prompt-masked, mean-pooled hidden states, ensuring that predictions are conditioned on the textual query rather than the full sequence. These heads serve as auxiliary supervision during training; the model's primary outputs remain autoregressive text.

Appendix B. Auxiliary-Head Ablation: Effect of Co-Training on Generative Performance

The auxiliary heads described in Section 3.1.1 could plausibly help in two distinct ways: by providing an additional, non-generative prediction pathway that can be read out at inference time, or by enriching the shared representation so that the generative pathway itself becomes more accurate. This appendix isolates the second effect.

We compare two configurations of CARE-X that differ only in their training objective:

- **Generative-only.** The backbone is trained with cross-entropy on the generative pathway alone; no auxiliary heads are attached.

- **Co-trained.** The identical backbone is trained with cross-entropy on the generative pathway jointly with focal loss on the classification heads and the composite spatial loss on the grounding head.

Crucially, **both configurations are evaluated in generative (autoregressive) decoding mode**; the auxiliary heads are not used to produce the reported numbers in either column. Any difference is therefore attributable to the effect of auxiliary supervision on the shared representation, not to substituting a discriminative readout for generation. The co-trained column corresponds to the CARE-X generative rows reported in Tables 5 and 4.

Table 9: Effect of auxiliary co-training on **generative** abnormality classification (Chest ImaGenome closed VQA). Both columns are evaluated by autoregressive decoding; the auxiliary heads are not read out. All metrics are higher-is-better (↑).

Metric	CARE-X (generative-only)	CARE-X (co-trained)	Δ
F1 ↑	0.895	**0.913**	+1.8 pp
Sensitivity (Recall) ↑	0.895	**0.932**	+3.7 pp
PPV (Precision) ↑	0.894	**0.895**	+0.1 pp

Table 10: Effect of auxiliary co-training on **generative** spatial grounding across four benchmarks. Both rows are evaluated by autoregressive decoding of normalized box coordinates; the detection head is not read out. All metrics are higher-is-better (↑).

Training	Anatomy (Chest ImaGenome)		Abnormality (VinDR)		Phrase (Padchest)		Phrase (MS)	
	mAP ↑	mIoU ↑	mAP ↑	mIoU ↑	mAP ↑	mIoU ↑	mAP ↑	mIoU ↑
CARE-X (generative-only)	0.526	0.473	0.279	0.206	0.367	0.257	0.424	0.312
CARE-X (co-trained)	**0.808**	**0.535**	**0.323**	**0.233**	**0.613**	**0.398**	**0.762**	**0.494**
Δ	+28.2 pp	+6.2 pp	+4.4 pp	+2.7 pp	+24.6 pp	+14.1 pp	+33.8 pp	+18.2 pp

Co-training improves the generative pathway on both task families. Classification F1 rises by +1.8 pp with a +3.7 pp sensitivity gain at essentially unchanged precision (Table 9), and every one of the eight grounding metrics improves, by between +2.7 pp and +33.8 pp (Table 10). Because the heads are inactive at evaluation time in both configurations, these gains cannot be explained by the presence of an extra prediction pathway. The auxiliary objectives instead act as structured supervision on the shared representation, and that benefit is inherited by autoregressive decoding—supporting the claim that discriminative supervision and generation reinforce one another rather than compete for capacity.

Appendix C. SFT Training Configuration

SFT. In the SFT stage, we initially train the model for one epoch with a learning rate of 1e-4 where the vision language adapter and the auxiliary heads are trainable. Then the model is trained for three epochs with parameter-efficient fine-tuning via LoRA (rank

r=32, α=32) and a learning rate of 5e-4 where the vision language adapter, the auxiliary heads and the LoRA adapters are trainable. In both stages the batch size is maintained at 128. In all stages we use AdamW optimizer with cosine learning rate scheduling and linear warm-up. For tasks with auxiliary heads, additional structured losses are applied:

- **Abnormality presence and tubes/lines placement:** The classification heads are trained with focal loss ($\alpha = 0.55$, $\gamma = 2.0$) to handle class imbalance between normal and abnormal samples.
- **Visual grounding:** The detection head is trained with a composite loss combining Generalized IoU loss ($\mathcal{L}_{\text{GIoU}}$, weight 5.0), mean IoU loss ($\mathcal{L}_{\text{mIoU}}$, weight 0.8), L1 coordinate regression loss ($\mathcal{L}_{\text{L1}}$, weight 7.0), and focal loss on box confidence scores ($\alpha = 0.55$, $\gamma = 2.0$, weight 2.0). Predicted boxes are matched to ground-truth boxes using the Hungarian algorithm on a cost matrix combining L1 distance, GIoU, mIoU, and confidence scores.

Appendix D. Dataset Details

We curate an instruction tuning dataset comprising approximately 5 million instances from multiple publicly available chest X-ray datasets, spanning diverse forms of supervision including free-text reports, categorical labels and bounding box annotations. Each instance pairs a frontal chest X-ray image (with prior images when available) with a user–assistant interaction for the target task. The full train/test distribution is provided in Table 11.

Report generation. For findings and impression generation, we draw from Interpret-CXR (Xu et al., 2024), which integrates MIMIC-CXR (Johnson et al., 2019), CheXpert (Irvin et al., 2019), PadChest (Bustos et al., 2020), BIMCV-COVID19 (Vayá et al., 2020), and OpenI (Demner-Fushman et al., 2016), alongside CheXpert-Plus (Chambon et al., 2024) and ReXGradient (Zhang et al., 2025a), yielding approximately 366K findings-generation and 686K impression-generation training samples.

Visual question answering. Clinical VQA tasks address the assessment of presence, negation, location, and classification of abnormalities as well as tubes and lines, along with differential diagnosis and geometric information extraction. Abnormality presence training draws from Chest ImaGenome (Wu et al., 2021), ReXVQA (Pal et al., 2025), NIH-CXR (Wang et al., 2017), SIIM (Abedalla et al., 2020), Candid-PTX (Feng et al., 2021), and CXR-LT (Holste et al., 2023) (∼1.2M samples). Abnormality location combines Chest ImaGenome and ReXVQA (∼730K), while abnormality classification uses Chest ImaGenome (236K). Negation assessment leverages ReXVQA (∼158K), differential diagnosis (∼46K) and geometric information assessment (∼2.2K) also draws from ReXVQA.

Tubes and lines tasks encompass multiple subtasks: presence detection (702K samples from Chest ImaGenome), which determines whether any tubes or lines are present; classification (∼42K samples from Interpret-CXR, CheXpert-Plus, and ReXGradient), which identifies the specific types of tubes, lines, or devices in a chest X-ray; placement description (∼199K samples from the same three report datasets), which characterizes their positioning; and abnormal placement detection (∼98K samples from Interpret-CXR, ReXGradient, CheXpert-Plus, and RANZCR (Seah et al., 2020)), which identifies presence of incorrectly positioned tubes, lines, or devices.

Visual grounding. Anatomical grounding uses 166K samples from Chest ImaGenome covering 29 lung anatomical regions. Phrase grounding combines MS-CXR and PadChest-GR (∼5.2K), while abnormality grounding additionally incorporates VinDR-CXR (∼21K total). For detailed dataset count refer to Table 11.

Table 11: Instruction dataset overview. Tasks marked with * employ auxiliary heads (classification or detection) in addition to the generative objective.

Task	**Training Datasets**	**Samples**	**Test Datasets**	**Samples**
Findings Generation	Interpret-CXR	217,268	MIMIC-CXR	2,345
	CheXpert-Plus	49,600	CheXpert-Plus	62
	ReXGradient	99,425	ReXGradient	6,271
Impression Generation	Interpret-CXR	396,111	MIMIC-CXR	2,345
	CheXpert-Plus	190,747	CheXpert-Plus	200
	ReXGradient	99,425	ReXGradient	6,271
Abnormality Classification	Chest ImaGenome	236,414	Chest ImaGenome	3,386
Abnormality Presence*	Chest ImaGenome	782,085	Chest ImaGenome	19,606
	ReXVQA	180,219	ReXVQA	14,602
	NIH-CXR	89,008	NIH-CXR	49,976
	SIIM	10,712	SIIM	1,377
	Candid-PTX	12,476	Candid-PTX	4,158
	CXR-LT	173,955		
Negation Assessment	ReXVQA	158,131	ReXVQA	14,918
Abnormality Location	Chest ImaGenome	700,000	Chest ImaGenome	23,785
	ReXVQA	30,382	ReXVQA	2,394
Geometric Information Assessment	ReXVQA	2,213	ReXVQA	168
Differential Diagnosis	ReXVQA	46,358	ReXVQA	168
Tubes & Lines Presence	Chest ImaGenome	702,181	Chest ImaGenome	11,145
Tubes & Lines Classification	Interpret-CXR	19,406	Chest ImaGenome	11,145
	CheXpert-Plus	16,029		
	ReXGradient	6,361		
Tubes & Lines Placement Description	Interpret-CXR	138,726	Interpret-CXR	1,331
	CheXpert-Plus	32,278	CheXpert-Plus	33
	ReXGradient	27,757		
Anatomical Grounding*	Chest ImaGenome	166,496	Chest ImaGenome	47,388
Phrase Grounding*	MS-CXR	815	MS-CXR	176
	PadChest-GR	4,335	PadChest-GR	1,238
Abnormality Grounding*	MS-CXR	800	NIH-CXR	984
	PadChest-GR	4,499	PadChest-GR	1,279
	VinDR-CXR	16,089		
Tubes & Lines Abnormal Placement*	Interpret-CXR	47,384	RANZCR	560
	ReXGradient	5,030		
	CheXpert-Plus	19,838		
	RANZCR	25,368		

Appendix E. Tubes & Lines Abnormal Placement

This section describes the construction and evaluation of our tubes & lines analysis framework for chest radiographs, focusing on both device presence identification and placement assessment. We first outline the dataset curation process, leveraging large-scale public resources to derive structured supervision for abnormal placement detection. We then present

a two-stage inference pipeline that decomposes the task into device detection followed by placement classification, enabling fine-grained and scalable evaluation. Finally, we report performance across multiple datasets, highlighting challenges in balancing sensitivity and specificity and demonstrating the effectiveness of our proposed approach.

E.1. Tubes & Lines Dataset Creation

As shown in Table 11, we construct the tubes & lines abnormal placement task using publicly available datasets, including Interpret-CXR and ReXGradient. We restrict the data to studies with frontal chest X-rays and corresponding radiology reports. These reports are processed using GPT-5.1 (OpenAI, 2025) to extract device-related phrases and determine placement status (normal vs. abnormal). For the prompt refer to Table 12.

Table 12: Prompt for Tube, Line, and Device Extraction

Role: Radiology expert interpreting chest X-rays.
Task: From a radiology report, extract all tubes/lines/devices and classify placement as *Normal*, *Abnormal*, or *Not Mentioned*.
Devices: Include vascular lines, airway tubes, GI tubes, cardiac devices, orthopedic implants, drains, surgical clips, and any visible devices. Exclude removed devices. Use "Unknown device" if unnamed.
Placement: Normal=correct/stable/appropriate; Abnormal=malpositioned/displaced/kinked/etc.; Not Mentioned=placement not stated.
Rules: (1) If input empty → "N/A"; if no devices → "N/A".
(2) Remove report-language (e.g., "report states").
(3) Preserve all clinical details exactly; no additions.
(4) Output must be a clean and independent
(5) For Normal/Abnormal: brief reason; for Not Mentioned: reason = "".
Special Rules: - If a device lacks a specific name, use "Unknown device". - Exclude devices that are removed. - If no devices are mentioned, return "N/A".

E.2. Inference Pipeline

We adopt a two-stage inference pipeline to evaluate the presence and placement of tubes and lines in chest radiographs.

Stage 1: Tubes & Lines Presence Identification : Given an input chest radiograph, the model performs open-ended, report-style generation to identify all visible tubes, lines, and medical devices, as detailed in Table 21. The output is a set of detected device categories (e.g., chest tube, internal jugular (IJ) line, nasogastric tube), enabling flexible identification of multiple devices within a single study.

Stage 2: Tubes & Lines Placement Classification : For each detected device from Stage 1, a follow-up query is constructed to evaluate its placement, using the "Tubes/Lines Abnormal Placement Detection" prompt described in Table 21. The model then classifies the placement of each device as either normal or abnormal.

Case-level Aggregation : Final study-level labels are obtained by aggregating device-level predictions. Specifically, if any detected device is classified as abnormal, the entire

Table 13: Tubes & Lines Presence Table for NH Clinical Dataset — Study 1

Model	Sensitivity	Specificity
CheXOne	**1.00**	0.00
MedGemma	0.99	0.19
CARE-X-SFT	0.89	**0.86**

study is labeled as abnormal. Conversely, a study is labeled as normal only if all detected devices are classified as normal.

In Table 13, we report the performance of tube and device presence classification on the NH Clinical Dataset (Study 1), as described in Section 3.3. Both CheXOne and MedGemma exhibit a strong bias toward predicting the presence of tubes or devices, even in negative cases. This results in near-perfect sensitivity (approx. 1.00) but severely compromised specificity (approaching 0), indicating that the model is unable to handle negatives. In contrast, CARE-X-SFT achieves a more balanced trade-off between sensitivity and specificity, demonstrating improved reliability in distinguishing true positive and negative cases.

E.3. Tubes & Lines Abnormality Placement Classification

Table 14 reports classification performance on the RANZCR Seah et al. (2020) Tubes & Lines Placement dataset, which encompasses Endotracheal tube (ETT), Nasogastric tube (NGT), and Central Venous Catheter (CVC) placement categories.

Among the evaluated models, CheXOne demonstrates low sensitivity (0.180), indicating a limited ability to correctly identify positive cases and missing out on substantial proportion of abnormalities, thereby reflecting poor predictive performance. MedGemma by contrast, demonstrates higher sensitivity (0.614) but at the cost of reduced specificity (0.383).

The proposed CARE-X-SFT variants exhibit a more balanced trade-off between sensitivity and specificity. The CARE-X-SFT-Generative model achieves competitive performance across both metrics, offering a meaningful improvement in balance over the baseline models. The CARE-X-SFT-CLFHead model attains the highest sensitivity among all evaluated models (0.783) while maintaining a reasonable level of specificity (0.639), demonstrating improved detection of abnormal placements without a disproportionate increase in false positives. This result supports the benefit of incorporating a dedicated classification head within the proposed framework.

Table 14: Classification performance of models on the RANZCR Tubes & Lines Placement dataset. Best values per metric are shown in **bold**.

Model	**Sensitivity**	**Specificity**
MedGemma	0.614	0.383
CheXOne	0.180	**0.960**
CARE-X-SFT-Generative	0.658	0.622
CARE-X-SFT-CLFHead	**0.783**	0.639

Appendix F. NH Dataset Distribution

As described in Section 3.3, in this section we break down the distributions of the datasets between the Study-1 and Study-2 datasets. Table 15 details the composition of the Study 1 evaluation cohort, comprising 1,047 de-identified chest radiographs annotated for five binary conditions by qualified radiologists. The distribution of the evaluated pathologies is severely imbalanced and negative-dominated, with positive case counts ranging from just 27 (Pneumoperitoneum) to 54 (Pneumothorax); the Tubes & Lines Presence prerequisite task is the exception, with 869 of the 1,047 studies device-positive. The three rarest findings – Pneumoperitoneum (2.6%), Mediastinal Shift (2.9%), and Fracture (3.5%) – each constitute under 4% of the cohort, reflecting realistic ICU and inpatient prevalence for high-acuity pathologies where missed diagnoses carry significant clinical risk. Tubes & Lines Abnormal Placement is treated as a conditional sub-task, evaluated only on the 869 radiographs confirmed to contain at least one device, of which just 38 exhibit abnormal placement – a prevalence of 4.4% within that subset. Further, individual radiographs may contain between zero and three distinct tube or line devices, introducing additional intra-image complexity. This low-prevalence, negative-dominated, and structurally heterogeneous OOD setting provides a rigorous stress-test of the model's ability to detect rare, high-acuity findings under realistic clinical conditions.

All chest radiographs were sourced from a cardiac intensive care unit (ITU) within NH, a quaternary care hospital in India and retrieved from an in-house EMR and data lake. Radiological annotations were derived exclusively from clinical reports authored by senior radiologists and consultants, ensuring expert-level label quality. Each report was digitally signed and timestamped, capturing structured findings across cardiopulmonary, osseous, and soft tissue domains, alongside relevant clinical context such as post-operative status (e.g., sternotomy, lines and tubes position). No additional re-annotation or crowdsourcing was performed; ground truth labels were extracted directly from these verified clinical reports, preserving real-world diagnostic standards reflective of the Indian subcontinent adult cardiac ITU population.

Table 15: Study 1: ICU and inpatient evaluation cohort from NH. All 1,047 radiographs are annotated for all six conditions; prevalence ranges from 2.6% to 5.2%.

Condition	Positive	Negative	Total
Tubes & Lines Presence	869	178	1,047
Tubes & Lines Abnormal Placement	38	831	869
Pneumoperitoneum	27	1,020	1,047
Mediastinal Shift	30	1,017	1,047
Pneumothorax	54	993	1,047
Fracture	37	1,010	1,047

In Study-2 , Table 16 summarises the composition of the Study 2 evaluation cohort, comprising 122 de-identified chest radiographs spanning five enlargement and mediastinal conditions. Critically, all cases are confirmed positive, with ground truth established via CT imaging rather than radiologist consensus alone – a design choice that eliminates the

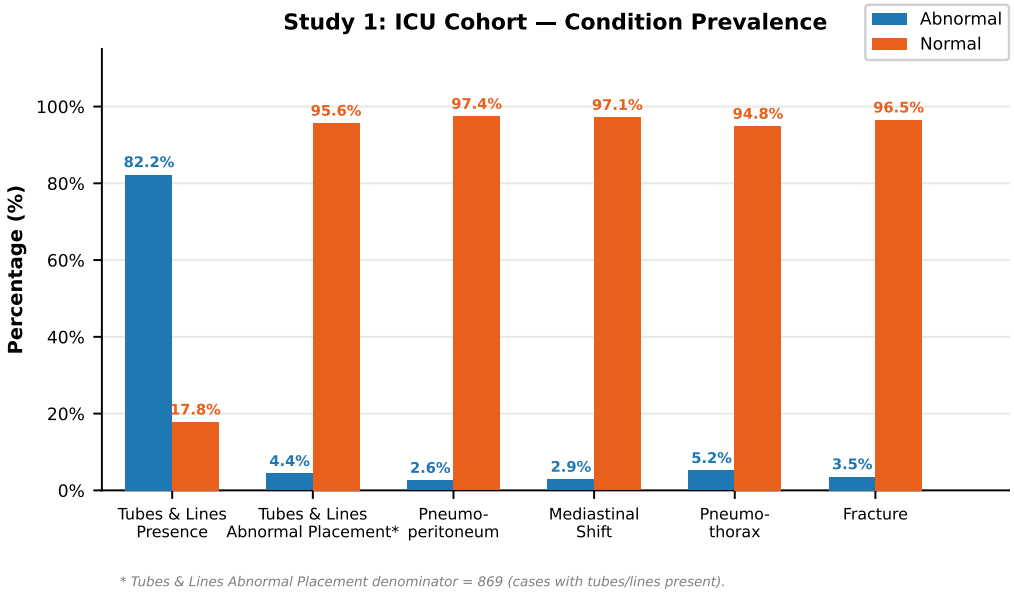


Figure 3: NH Dataset Prevalence Distribution for Inpatient & ICU Conditions

subjectivity inherent in borderline enlargement assessments on plain radiographs. The cohort is dominated by Aortic Enlargement (69.23%), consistent with its higher prevalence in outpatient referral populations, followed by Hilar Mass (16.92%). Other findings – Aortic Dissection (4.62%), Pulmonary Artery Enlargement (1.54%), and other mediastinal pathologies such as Mediastinal Widening and Cardiomegaly (7.69%) – collectively constitute the remainder. Since the cohort contains no true negatives by design, sensitivity (recall) serves as the sole meaningful performance metric.

Table 16: Study 2: Outpatient enlargement evaluation cohort from NH. All the pathologies are positive cases confirmed via CT Imaging.

Pathology	**Percentage (%)**
Aortic Enlargement	69.23
Hilar Mass	16.92
Aortic Dissection	4.62
Pulmonary Artery Enlargement	1.54
Others (Mediastinal Widening , Cardiomegaly , etc)	7.69
Total	**100.00**

F.1. Threshold-Selection Sensitivity Analysis

The per-condition decision thresholds reported in Table 7 were selected on the same 1,047 radiographs used for evaluation, which risks optimistic bias in the reported operating points. To quantify this effect, we repeated the analysis under 3-fold cross-validation with disjoint

splits: threshold sweeps were performed on the merged training folds, and the selected thresholds were then applied to the held-out test fold. Table 17 compares the resulting cross-validated metrics against the originally reported values for the four unconditional binary conditions. Tubes & Lines Abnormal Placement is excluded because it is a conditional subtask evaluated only on the 869 device-positive radiographs, of which just 38 exhibit abnormal placement—too few positives for stable threshold estimation under 3-fold splitting.

Table 17: Cross-validated versus reported operating points for CARE-X-SFT. Differences are uniformly small ($|\Delta| \leq 0.038$).

	Sensitivity			Specificity		
Pathology	3-Fold CV	Reported	Δ	3-Fold CV	Reported	Δ
Fracture	0.622	0.620	+0.002	0.634	0.640	−0.006
Pneumothorax	0.833	0.830	+0.003	0.753	0.750	+0.003
Pneumoperitoneum	0.852	0.890	−0.038	0.948	0.940	+0.008
Mediastinal Shift	0.833	0.830	+0.003	0.855	0.860	−0.005

Differences across all four conditions are uniformly small ($|\Delta| \leq 0.038$), with the largest deviation occurring on Pneumoperitoneum, the rarest finding in the cohort. This indicates that selecting thresholds on the evaluation set had negligible practical impact on the reported performance, and that the chosen operating points remain stable across unseen folds.

Appendix G. Likelihood Ratio Analysis of CARE-X Under Distribution Shift on Rare ICU Conditions

Under OOD evaluation, CARE-X-SFT-CLFHead demonstrated consistently meaningful likelihood ratios (LR) across all five conditions, a property neither CheXOne nor MedGemma achieved.

In a cardiac ITU population where pathologies such as pneumoperitoneum and mediastinal shift carry high mortality risk if missed, a model must perform reliably across the full diagnostic spectrum. CARE-X-SFT-CLFHead achieved an LR^+ of 14.83 for pneumoperitoneum and 5.93 for mediastinal shift (Table 18), indicating clinically meaningful upward revision of post-test probability, while maintaining LR^- values of 0.12 and 0.20 respectively, approaching rule-out utility. This bidirectional discriminability is essential in triage-assist settings where both false negatives and false positives carry patient safety consequences.

The severe class imbalance in rare ITU conditions means a model can achieve deceptively high specificity by defaulting to negative predictions. MedGemma illustrates this on pneumoperitoneum, where Sensitivity = 0.00 and Specificity = 1.00 yields a degenerate LR with no diagnostic value.

These findings suggest CARE-X-SFT-CLFHead is the most promising candidate for clinical decision support in this setting—not because it achieves the highest individual metric on any single condition, but because it is the only model that consistently shifts post-test probability in a diagnostically actionable direction across conditions encountered in an adult cardiac ITU, maintaining this property under distribution shift.

Table 18: Positive and Negative Likelihood Ratios for Inpatient and ICU Pathology Classification Under Out-of-Distribution Performance Evaluation. Best LR^+ (highest) and best LR^- (lowest) per condition are **bolded**, excluding degenerate values. †Decision thresholds tuned per condition. *Degenerate value: Specificity = 1.00 yields undefined LR^+ (division by zero). ‡Degenerate case: Sensitivity = 0.00 and Specificity = 1.00 simultaneously; both LRs are undefined. $LR^+ > 10$ indicates strong rule-in; $LR^- < 0.1$ indicates strong rule-out.

	Fracture		**Med. Shift**		**Pneumop.**		**Pneumotx.**		**Tubes & Lines**	
Model	**LR^+**	**LR^-**	**LR^+**	**LR^-**	**LR^+**	**LR^-**	**LR^+**	**LR^-**	**LR^+**	**LR^-**
CheXOne	**4.10**	0.66	3.64	0.26	**33.50**	0.34	3.04	**0.21**	1.00	1.00
MedGemma	∞^*	0.95	2.13	0.00	–‡	–‡	1.93	0.66	1.38	0.94
CARE-X-SFT-CLFHead†	1.72	**0.59**	**5.93**	**0.20**	14.83	**0.12**	**3.32**	0.23	**2.87**	**0.44**

Appendix H. Evaluation Prompts

We use standardized prompt templates across report generation, grounding, and VQA tasks to ensure consistent evaluation, covering findings/impression generation, spatial grounding, and clinically relevant question answering. The prompts used are mentioned in Tables 19, 20, and 21.

Table 19: Prompt templates for findings and impression generation tasks.

Task	Task Setting	Input Prompt
Report Generation	Findings Generation	Given the chest X-rays and the indication section, write the findings.
	Impression Generation	Given the chest X-rays and the indication section, write the impression.

Table 20: Prompt templates for visual grounding tasks.

Task Setting	Input Prompt
Visual Grounding	Ground the location of {label / phrase / region} in the chest X-ray.

Table 21: Prompt templates for VQA tasks.

Task	Task Setting	Chestimagenome Prompt	ReXVQA Prompt
VQA	Abnormality Presence (image-level)	Is there evidence of any abnormalities?	Does the image show any abnormalities?
	Abnormality Presence (by-finding)	Is <> seen in the image?	Is <> present in the image?
	Abnormality Location	In which location is the <> seen?	Where can the <> be observed?
	Abnormality Classes	List the abnormalities seen in the image.	What findings can be identified in this image?
	Tubes/Lines Presence (image-level)	Is there any presence of tubes, lines, or devices visible in the chest X-ray?	Are there any medical devices present in this chest X-ray?
	Tubes/Lines Presence (tube-type)	Is there any indication of an <> in this chest X-ray?	Is a <> visible in this chest X-ray?
	Tubes/Lines Location		Where is the <> located in the chest X-ray?
	Tubes/Lines Type	What kinds of tubes, lines or devices are visible in this image of a chest x-ray?	Which types of medical devices are visible?
	Tubes/Lines Placement Description	Describe the placement of <>	Where are the tubes and lines located in the chest X-ray?
	Tubes/Lines Abnormal Placement Detection	Is the placement of <> abnormal?	

Appendix I. DAPO Training Details

I.1. Hyperparameters

Table 22 reports the full DAPO training configuration used across all tasks.

Table 22: DAPO training hyperparameters. All experiments use LoRA-based parameter-efficient fine-tuning with a single training epoch on 8 NVIDIA GPUs.

Parameter	Value
Epochs	1
Batch size	8
Generations per prompt	8
Learning rate	5×10^{-6}
Max completion length	1024
Temperature	0.7
Gradient accumulation steps	1
Max gradient norm	1.0
Precision	bf16
LoRA Configuration	
Rank (r)	16
Alpha (α)	32
Dropout	0.05
Target modules	`qkv_proj`, `o_proj`, `gate_up_proj`, `down_proj`

I.2. Report Generation Reward Curriculum

Jointly optimizing all three report generation rewards from initialization risks reward hacking, where the model exploits inter-metric correlations rather than improving genuine clinical quality. We therefore employ a three-stage reward curriculum that progressively introduces reward complexity:

1. **Stage 1** (5,000 steps): BERTScore (Zhang et al., 2020) and RadGraph (Delbrouck et al., 2022) with equal weights ($w_{\text{BERT}} = 0.5$, $w_{\text{RG}} = 0.5$), establishing semantic and entity-level fidelity.

2. **Stage 2** (6,000 steps): GREEN (Ostmeier et al., 2024) is introduced with increased weight ($w_{\text{BERT}} = 0.25$, $w_{\text{RG}} = 0.25$, $w_{\text{GREEN}} = 0.5$), shifting emphasis toward holistic clinical completeness.

3. **Stage 3** (7,500 steps): Returns to BERTScore and RadGraph with equal weights ($w_{\text{BERT}} = 0.5$, $w_{\text{RG}} = 0.5$) for final stabilization.

This curriculum first establishes semantic and entity-level fidelity through BERTScore and RadGraph, then shifts emphasis to holistic clinical completeness via GREEN, before returning to the foundational rewards for final stabilization. For closed VQA and spatial grounding, the reward signals remain fixed throughout training (binary reward for VQA; mIoU, gIoU, and box count for grounding).

I.3. Reward Model Implementation

Table 23 summarizes the models used for reward computation during DAPO training. All reward models are served via FastAPI for online computation during rollouts.

Table 23: Reward model implementation details for DAPO training.

Metric	Model	Key Configuration
BERTScore	`microsoft/deberta-xlarge-mnli`	No baseline rescaling; batch 64
RadGraph	`radgraph-xl`	Partial reward level (entity + relation)
GREEN	`StanfordAIMI/GREEN-radllama2-7b`	LLaMA-2 7B fine-tuned for error detection

Appendix J. Residual Gaps in Reward-Aligned Learning

This appendix provides detailed analysis of the three settings where CARE-X-RL does not achieve a clean win, corresponding to the limitations summarised in Section 5.1.

J.1. Report Generation: Competitor Leads on CheXpert-Plus and IU-Xray GREEN

Where competitors lead, dataset-specific reporting conventions offer a plausible explanation. CheXOne-R1 achieves higher BERTScore, SEmb, and 1/RadCliQ-v1 on CheXpert-Plus, whose reports tend to be more concise and structured than MIMIC-CXR or ReXGradient. CheXOne-R1's reward configuration may be better aligned with this reporting style. MedGemma achieves the highest GREEN on IU-Xray and ReXGradient, indicating particular strength in holistic report quality on these benchmarks. These patterns suggest that reward alignment effects are partially dataset-dependent, motivating future work on adaptive reward weighting across reporting conventions.

J.2. Closed VQA: Binary Reward Fails Geometric Reasoning

Geometric reasoning is the only category where DAPO underperforms SFT. Geometric information assessment drops by -1.1 pp, the only category-level decrease. Spatial reasoning questions often admit partially correct answers (e.g., approximate anatomical descriptions) that contain useful clinical information but receive zero reward under binary-reward scoring. This limitation suggests that graduated reward functions—awarding partial credit for spatially approximate answers—may be needed to improve quantitative spatial reasoning without sacrificing gains elsewhere.

J.3. Spatial Grounding: Abnormality Localization on VinDR

Abnormality grounding on VinDR remains the most challenging setting. On VinDR, all our model variants trail RadVLM in mAP. While DAPO improves over SFT generative (+21.7%), the gap to RadVLM—which benefits from grounding-specific pretraining—persists. Abnormality localization involves small, variably shaped findings that may be

entirely absent, making it particularly sensitive to fine-grained spatial representations. Targeted data augmentation or abnormality-specific reward shaping may be needed to close this gap.

Appendix K. Quantitative Reasoning Details

This appendix provides supporting details for the tool-augmented quantitative reasoning pipeline. We first present the full specification of clinical conditions, anatomical structures, measurement tools, and diagnostic thresholds, followed by dataset construction details including source repositories, anatomy overlay generation, and the structured evaluation format (Appendix K.1). We then report a view-classification ablation for cardiomegaly and mediastinal widening and perception-only baselines across all conditions. Appendix K.3 analyses why the quantitative reasoning gap is a tool-access problem rather than a model-capacity limitation, and Appendix K.4 presents an error analysis of the two principal failure modes. Finally, we include representative positive and negative measurement visualisations for each condition.

Table 24: Clinical conditions, segmented anatomical structures, measurement tools, diagnostic thresholds, and corresponding measurement metrics. CM and MW use view-dependent thresholds (AP vs. PA); AK, AAE, and DA use expert-defined ratio-based criteria.

Condition	Abbr.	Structures	Tools	Thresholds Used	Metric
Cardiomegaly	CM	H, RL, LL	`measure_cardiac_width`, `measure_thoracic_width`, `compute_ctr`	PA: CTR ≥ 0.50; AP: CTR ≥ 0.55 $\Rightarrow$ Cardiomegaly present	$\text{CTR} = \frac{W_{\text{heart}}}{W_{\text{lung}}}$
Mediastinal Widening	MW	UM, RL, LL	`measure_mediastinal_width`, `measure_thoracic_width`, `compute_mtr`	PA: MCR ≥ 0.25; AP: MCR ≥ 0.33 $\Rightarrow$ Widening present	$\text{MCR} = \frac{W_{\text{medi}}}{W_{\text{lung}}}$
Aortic Knob Enl.	AK	AA, DA, Tr, TB	`find_aortic_knob_edges`, `measure_trachea_width`, `compute_ak_ratio`	$R \geq 2.5$ $\Rightarrow$ Enlargement present	$R = \frac{W_{\text{knob}}}{W_{\text{trachea}}}$
Asc. Aorta Enl.	AAE	AsA, H, Tr, TB	`find_asc_aorta_edges`, `measure_trachea_width`, `compute_aae_ratio`	$0.1 \leq r < 0.3$ $\Rightarrow$ Enlargement present; $r = 0 \Rightarrow$ Not present	$r = \frac{A_{\text{ext}}}{A_{\text{total}}}$
Desc. Aorta Enl.	DA	DA, Tr, TB	`find_desc_aorta_edges`, `measure_trachea_width`, `compute_da_ratio`	$R \geq 2.5$ $\Rightarrow$ Enlargement present	$R = \frac{W_{\text{desc}}}{W_{\text{trachea}}}$

H = Heart; RL = Right Lung; LL = Left Lung; UM = Upper Mediastinum; AA = Aortic Arch; AsA = Ascending Aorta; DA = Descending Aorta; Tr = Trachea; TB = Trachea Bifurcation.

Metric notation: W denotes width measurements extracted from anatomical landmarks; A denotes area measurements computed from segmented contours;

K.1. Dataset Details

Our evaluation dataset is drawn from three publicly available chest X-ray repositories. **MIMIC-CXR** (Johnson et al., 2019) provides the largest share of samples across all five conditions, contributing frontal radiographs with associated radiology reports and metadata including AP/PA view labels. **PAXRay** (Seibold et al., 2022) supplements samples

across all five conditions. **VinDr-CXR** (Nguyen et al., 2022) contributes samples for three conditions—aortic knob enlargement, ascending aorta enlargement, and descending aorta enlargement. The per-condition distribution across datasets is detailed in Table 25; in all experiments, a single frontal radiograph per sample serves as the primary image input.

Anatomy overlay images. In addition to the raw frontal radiographs, we generate anatomy overlay images for each sample by superimposing condition-relevant segmentation masks onto the original chest X-ray. For a given condition, only the anatomical structures required for its diagnostic measurement (as listed in Table 24) are overlaid—for example, a cardiomegaly sample includes heart, right lung, and left lung masks. These overlays were produced using **CXAS** model, which provides pixel-level anatomical segmentations for frontal chest radiographs. The overlay images serve as supplementary visual input, enabling the VLM to localise relevant structures more precisely during tool-augmented inference.

Quantitative Reasoning Data Construction For evaluation with Qwen-based vision–language model, we construct structured test files tailored to both perception-only and tool-augmented (measurement) settings. Each test instance contains a `current_image`, corresponding to the original frontal chest X-ray, and a `spl_image`, the anatomy overlay image described above. In perception-only test files, the model is provided with these images along with a system prompt and a natural-language diagnostic query, and performance is assessed based on the final textual prediction. In contrast, measurement test files additionally include an explicit schema of available tools for anatomical measurements (Table 24), requiring the output of one tool as input to another (e.g., cardiac and thoracic width are prerequisites for `compute_ctr`). The ground-truth labels are determined from tool-based measurements using view-specific diagnostic thresholds (e.g., CTR ≥ 0.55 for AP views in cardiomegaly).

K.2. CXReasonBench and CheXStruct.

Our evaluation pipeline builds on the CheXStruct framework from CXReasonBench (Lee et al., 2025) in two ways. First, we use its segmentation pipeline to generate pixel-level anatomical masks for each frontal chest radiograph, which serve as the basis for both the anatomy overlay images used in the dual-image setting and the coordinate inputs consumed by the measurement tools. Second, we adopt its condition-specific assessment methods—including geometric measurement procedures and diagnostic thresholds—to define the five pathology conditions evaluated in this work (Table 24). Ground-truth labels for each sample are derived by applying these thresholds to the CheXStruct-computed measurements.

Table 25: Quantitative Reasoning Data Distribution

Pathology	Dataset	Pos	Neg	Total
Cardiomegaly	MIMIC	1140	647	1787
	PAXRAY	233	155	388
	Total	1373	802	**2175**
Mediastinal Widening	MIMIC	827	551	1378
	PAXRAY	315	210	525
	Total	1142	761	**1903**
Aortic Knob Enl.	MIMIC	200	133	333
	PAXRAY	123	82	205
	VINDR	87	87	174
	Total	410	302	**712**
Asc. Aorta Enl.	MIMIC	57	38	95
	PAXRAY	45	30	75
	VINDR	25	25	50
	Total	127	93	**220**
Desc. Aorta Enl.[†]	MIMIC	1	0	1
	PAXRAY	3	2	5
	VINDR	1	0	1
	Total	5	2	**7**

[†]Limited public dataset availability for descending aorta enlargement resulted in only 7 test samples; results should be interpreted with caution.

Table 26: View-classification ablation for Cardiomegaly (CM) and Mediastinal Widening (MW) (%). All settings use tool-augmented measurement; they differ in how the AP/PA view label is obtained. **Bold** indicates the best value per metric.

	(A) CM			(B) MW		
Setting	Sens	Spec	F1	Sens	Spec	F1
Meas. + Perception	**96.94**	79.05	92.69	**96.85**	74.77	90.66
Meas. + GT View	96.07	**93.02**	**96.00**	95.42	**99.52**	**97.47**

Table 27: Perception-only classification performance (%)—the model classifies abnormality directly from the image without any measurement tools. A checkmark (✓) in the Overlay column indicates the anatomy-overlay image was provided (2-image setting).

			Classification Metrics (%)		
Condition	Model	Overlay	Sens	Spec	F1
CM	Qwen		66.21	62.84	70.47
	Qwen	✓	86.96	20.60	74.56
	CheXOne		85.00	68.33	83.54
	MedGemma		97.23	24.81	80.64
	CARE-X-SFT		87.28	79.30	86.86
	CARE-X-SFT	✓	92.33	88.03	92.22
MW	Qwen		26.62	83.84	38.75
	Qwen	✓	88.18	18.00	72.63
	CheXOne		70.58	60.05	71.58
	MedGemma		57.62	62.81	63.18
	CARE-X-SFT		88.09	72.67	85.40
	CARE-X-SFT	✓	94.40	87.12	93.01
AK	Qwen		16.10	89.74	26.04
	Qwen	✓	61.71	41.72	60.31
	CheXOne		77.07	48.34	71.66
	MedGemma		27.07	86.42	39.50
	CARE-X-SFT		82.93	70.20	80.95
	CARE-X-SFT	✓	88.29	80.13	87.02
AAE	Qwen		0.00	100.00	0.00
	Qwen	✓	27.56	82.80	39.33
	CheXOne		79.53	47.31	72.92
	MedGemma		11.81	94.62	20.41
	CARE-X-SFT		82.68	88.17	86.42
	CARE-X-SFT	✓	53.54	89.25	66.34
DA†	Qwen		0.00	100.00	0.00
	Qwen	✓	20.00	50.00	28.57
	CheXOne		100.00	100.00	100.00
	MedGemma		0.00	100.00	0.00
	CARE-X-SFT		100.00	100.00	100.00
	CARE-X-SFT	✓	100.00	100.00	100.00

†Limited public dataset availability for descending aorta enlargement resulted in only 7 test samples; results should be interpreted with caution.

K.3. Tool-Augmented Inference Analysis

The quantitative reasoning gap is a tool-access problem, not a model-capacity problem. We benchmark four VLMs under identical perception-only conditions on the five measurement-dependent tasks (CM, MW, AK, AAE, DA), evaluated by F1: Qwen3-VL-4B-Instruct, CheXOne (Qwen2.5VL-3B), MedGemma (medgemma-1.5-4b-it), and CARE-X-SFT (Table 27). No externally-released baseline—general-purpose (Qwen3-VL) or medical (CheXOne, MedGemma)—exceeds 84% F1 on CM, MW, AK, or AAE, and **this ceiling holds regardless of medical specialisation—domain pretraining is not the axis of variation on these ratio-based tasks.** CARE-X-SFT is the strongest perception-only configuration, yet even with the anatomy overlay it peaks at 93.0% F1 (MW) and remains below tool-augmented inference on every condition except DA, where both reach 100% F1 on only 7 samples. What moves the metric is precise measurement, not specialisation: adding deterministic tools to the base Qwen3-VL matches or exceeds every perception-only medical VLM without task-specific training, so **tool augmentation is the lever.** This comparison is scoped to these five tasks and to F1; by holding the model family fixed (Qwen3-VL with vs. without tools) it isolates tool access and is **not** a domain-vs-general claim. Medical pretraining remains decisive for the perception-driven tasks that dominate clinical reporting—report generation, VQA, grounding, and rare-ICU classification (Tables 2–4, 7).

K.4. Residual failure modes under tool-augmented inference.

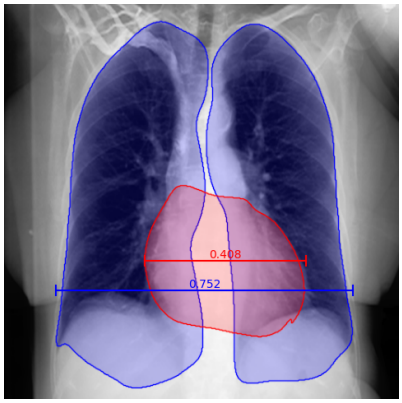
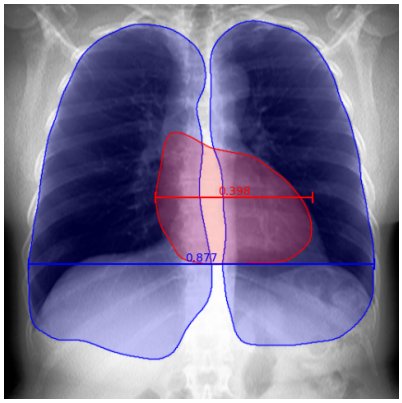
Despite the strong overall performance of tool-augmented inference, a detailed examination of failure cases reveals two distinct error modes.

(i) Numerical comparison errors near decision boundaries. When a computed measurement falls close to the classification threshold, the VLM occasionally misclassifies the comparison direction—for example, asserting CTR $\geq$ 0.55 when the measured value is 0.51, yielding a false-positive cardiomegaly diagnosis. This failure mode is most pronounced for conditions with tight threshold margins: cardiomegaly's AP/PA gap of only 0.05 leaves minimal room for imprecise reasoning, and mediastinal widening is similarly affected. In contrast, conditions with wider boundaries—such as AK and AAE (ratio threshold ≥ 2.5)—achieve >99% F1 (Table 26 and Table 6), as measurement values typically fall well above or below the threshold. Errors thus concentrate in a narrow band around the decision boundary rather than occurring uniformly across the measurement range.

(ii) Hedging bias in perception-only inference. Without access to measurement tools, VLMs resort to qualitative visual assessment, introducing a systematic hedging bias that manifests as extreme sensitivity–specificity imbalance (Table 27). The direction varies by model: Qwen achieves high sensitivity but near-zero specificity on CM (2-image), defaulting to "abnormal" when uncertain, while MedGemma inverts this pattern on AK, defaulting to "normal" and missing true abnormalities. This divergence reflects model-specific priors learned during pretraining rather than properties of the conditions themselves, and carries direct clinical consequences—low specificity inflates false-positive rates, while low sensitivity (poor NPV) translates to missed abnormalities in screening.

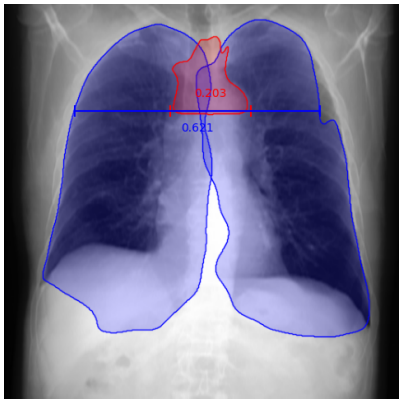
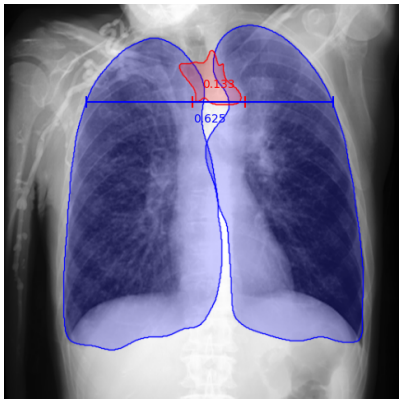
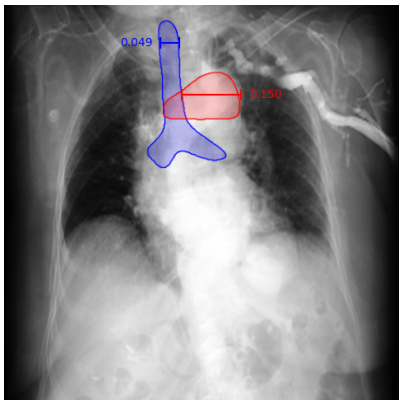
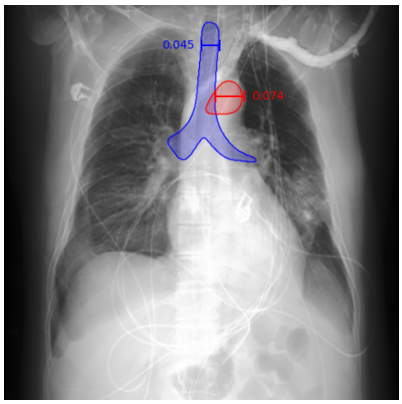
Summary. Both failure modes point to a common root cause: LLM-intrinsic limitations in numerical reasoning. The measurement tools produce accurate values; errors arise when the VLM must interpret and compare those values against thresholds, or when it lacks quantitative evidence entirely. This attribution is supported by the observation that tool augmentation eliminates hedging bias almost entirely (specificity improves by 55.9 pp on average) and that residual errors cluster exclusively near decision boundaries.

Table 28: Geometric measurements for positive and negative samples across pathologies. Each cell shows the visualization image with key measurement values below.

Pathology	**Positive Sample**	**Negative Sample**
Cardiomegaly (CM)		
	CTR = 0.54 Heart Width$_n$ = 0.408 Lung Width$_n$ = 0.752 View = PA	CTR = 0.45 Heart Width$_n$ = 0.398 Lung Width$_n$ = 0.877 View = AP

Cardiomegaly (CM): Enlarged heart silhouette on chest X-ray, diagnosed by the cardiothoracic ratio (CTR), defined as the maximal horizontal cardiac diameter divided by the maximal horizontal thoracic diameter. A CTR exceeding 0.50 in PA view indicates cardiomegaly.

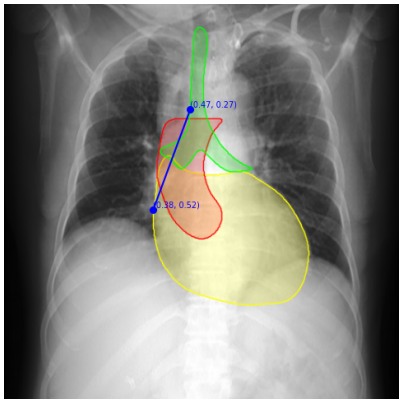
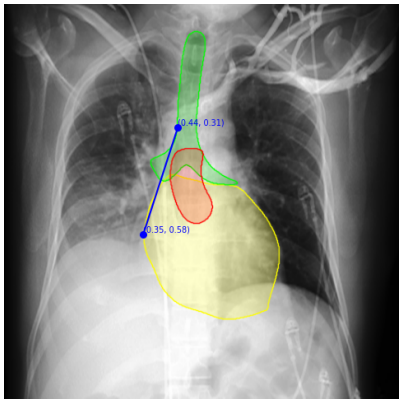
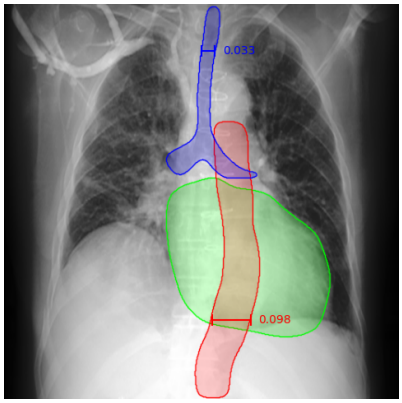
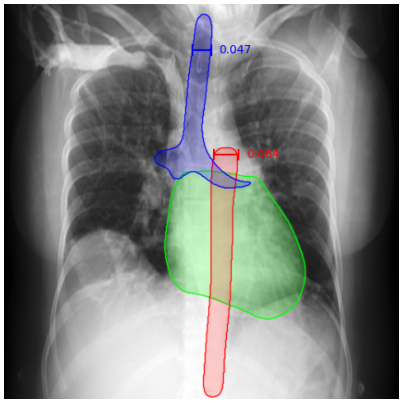
(continued from previous page)

Pathology	Positive Sample	Negative Sample
Mediastinal Widening (MW)		
	MCR = 0.33 Medi. Width_n = 0.203 Lung Width_n = 0.621 View = PA	MCR = 0.21 Medi. Width_n = 0.133 Lung Width_n = 0.625 View = AP
Aortic Knob Enlargement (AKE)		
	Ratio = 3.067 AK Width_n = 0.150 Trachea Width_n = 0.049	Ratio = 1.640 AK Width_n = 0.074 Trachea Width_n = 0.045

Mediastinal Widening (MW): Abnormal broadening of the mediastinum, traditionally defined as a mediastinal width >8 cm at the aortic arch level on PA chest X-rays. Since absolute measurements are not feasible in image-only settings, it is assessed here as the ratio of mediastinal width to thoracic width at the same level.

Aortic Knob Enlargement (AKE): Prominence of the aortic arch along the left mediastinal border. Quantified as the ratio of the maximum aortic knob width to the median tracheal width, using the trachea as a stable anatomical reference for reproducible assessment.

(continued from previous page)

Pathology	Positive Sample	Negative Sample
Ascending Aorta Enlargement (AAE)		
	Ratio = 0.175 Total Area$_n$ = 0.038 Extension Area$_n$ = 0.007	Ratio = 0.000 Total Area$_n$ = 0.016 Extension Area$_n$ = 0.000
Descending Aorta Enlargement (DAE)		
	Ratio = 2.941 DA Width$_n$ = 0.098 Trachea Width$_n$ = 0.033	Ratio = 1.375 DA Width$_n$ = 0.064 Trachea Width$_n$ = 0.047

Ascending Aorta Enlargement (AAE): Abnormal dilation of the ascending aorta, visible along the right mediastinal border. Defined by whether the aorta extends beyond an imaginary line connecting the right heart border and the inner margin of the right lung.

Descending Aorta Enlargement (DAE): Widening of the descending thoracic aorta, often seen as a prominent left paraspinal contour. Assessed via a ratio-based measurement between the aorta's maximum width and the median tracheal width, offering consistent evaluation.

Appendix L. Extended Related Work

This appendix provides a detailed survey of work most closely related to CARE-X across the three research threads it builds on: radiology vision-language models, reinforcement learning for radiology, and quantitative reasoning with tool use.

Radiology VLMs. Vision-language models for chest X-ray interpretation have advanced rapidly along two axes: task breadth and training signal. MedGemma (Sellergren et al., 2025) adapts a general-purpose VLM to medical imaging but retains a purely generative architecture without radiology-specific prediction heads. MedVersa (Zhou et al., 2026) goes further by coupling its LLM orchestrator with dedicated vision modules, though these modules are invoked as post-hoc specialists rather than co-trained with the generative objective. RadVLM (Deperrois et al., 2025) demonstrates that joint multitask SFT over report generation, classification, and grounding yields complementary gains across tasks, though it does not incorporate reinforcement learning. CheXOne (Zhang et al., 2026) represents the most comprehensive effort to date: a 3 B-parameter model trained on 14.7 M samples spanning 36 tasks with GRPO across VQA, report generation, and grounding, though the architecture remains purely generative. Rad-Phi4-Vision-CXR (Ranjit et al., 2025) introduced focal-loss classification and composite-loss grounding heads alongside a generative backbone, showing that structured supervision improves report fidelity; however, these auxiliary heads are trained independently and do not cross-train with the generative cross-entropy loss.

Reinforcement learning for radiology. UniRG-CXR (Liu et al., 2026) pairs SFT with GRPO on a Qwen3-VL-8B backbone, optimising a composite reward that aggregates lexical and clinical metrics; this yields state-of-the-art on the ReXrank benchmark but restricts RL to report generation. Gundersen et al. (Gundersen et al., 2025) extend GRPO to both report generation and visual grounding atop an updated RadVLM (Qwen3-VL), finding that RL improves both tasks while explicit chain-of-thought traces provide no additional benefit. CheXOne offers the broadest RL coverage (VQA, report generation, and grounding), yet initialises entirely from a generative SFT checkpoint without discriminative pre-training. CARE-X differs in two respects: DAPO is initialised from a model already strengthened by auxiliary-head co-training, and the resulting combination closes the gap between autoregressive spatial decoding and dedicated detection heads, a result no prior RL-only method has achieved.

Quantitative reasoning and tool use. Measurement-based diagnosis remains an open challenge for radiology VLMs. CXReasonBench (Lee et al., 2025) introduced CheXStruct, a structured diagnostic pipeline that derives intermediate reasoning steps (anatomical segmentation, landmark detection, measurement computation, and clinical-threshold application) and evaluated 12 VLMs across 12 diagnostic tasks, finding that even the strongest models fail to reliably link clinical knowledge with anatomically grounded measurements. CXReasonAgent (Lee et al., 2026) addresses this gap by coupling an LLM with CheXStruct's diagnostic tools, but the backbone receives only tool-derived features and never observes the image directly, precluding joint visual–quantitative reasoning. CARE-X integrates tool calling within a VLM that retains full visual access, enabling hybrid inference (e.g., perceiving view orientation visually while invoking measurement tools), a capability no prior system supports.