

Med-AR: Autoregressive Vision-Language Pretraining for Long-Tailed Chest X-Ray Classification and Uncertainty-Aware Evaluation

Janhavi Prabhu, Sahil, Akshay V, Shivam Shukla, Manoj Tadepalli, and
Preetham Putha

Qure.ai, Bangalore, India
 janhavi.prabhu@qure.ai
 sahil@qure.ai
 Akshay.V@qure.ai
 shivam.shukla@qure.ai
 manoj.tadepalli@qure.ai
 preetham.putha@qure.ai

Abstract. Long-tailed chest X-ray classification requires visual representations that capture both common abnormalities and subtle, infrequent findings. We propose **Med-AR-8B** and **Med-AR-2B**, two radiology-native autoregressive vision-language models pretrained with structured reports, abnormality-focused text, and region annotations. We evaluate the transfer of their visual encoders to multi-label classification against contrastive, self-supervised, and supervised pretrained encoders, including Med-CLIP, CheXFound, EVA-Base, ARK, and BioViL-T, using a common ML-Decoder classification head. To assess fine-grained recognition, we also construct LLM-expanded, report-derived label sets for MIMIC-CXR and CheXpert. Across PadChest, MIMIC-CXR, and CheXpert, Med-AR-8B outperforms Med-CLIP in mean AUROC and AUPRC for head, medium, and tail findings. On MIMIC-CXR, it increases tail-label mean AUPRC from 0.1033 to 0.1441. Med-AR-2B achieves the strongest discrimination results on PadChest. Across the broader encoder comparison, a Med-AR variant achieves the highest mean AUROC and AUPRC in every reported prevalence group on each public dataset. Both Med-AR variants also achieve lower excess area under the risk-coverage curve than Med-CLIP on all three public datasets, indicating improved selective-prediction performance under the evaluated protocol. Internal results are metric-dependent, with Med-CLIP retaining advantages in overall and tail AUPRC and in selective prediction. These findings establish Med-AR as a strong pretraining recipe for long-tailed chest X-ray classification on the evaluated public benchmarks and demonstrate the value of assessing discrimination and selective prediction together.

Keywords: MedCLIP · Vision-Language Models · Long-Tailed Classification · Uncertainty · Autoregressive Pretraining · Chest X-ray

1 Introduction

Chest X-rays (CXRs) are among the most frequently performed imaging studies for screening, triage, and longitudinal monitoring of thoracic disease, making robust automated abnormality recognition clinically important. However, multi-label CXR classification remains challenging because thoracic findings exhibit a highly long-tailed distribution: a small subset of common abnormalities accounts for the majority of positive samples, while many clinically significant findings appear only infrequently [16,22]. Consequently, models may achieve strong aggregate metrics yet fail on rare conditions, where reliable performance is often most clinically necessary. This problem is further complicated by disease co-occurrence, reporting variability, label noise, and dataset heterogeneity.

Rare abnormality recognition in CXRs also differs from standard image-level classification. Many uncommon or clinically subtle findings are spatially localized, low contrast, or interpretable only in relation to surrounding anatomical structures. Although broad abnormalities such as cardiomegaly or large pleural effusions can often be identified from coarse global structure, other findings require fine-grained local evidence together with anatomical and contextual understanding. Effective long-tail CXR classification therefore depends on representations that preserve localized visual detail while simultaneously maintaining clinically meaningful global context, rather than relying solely on pooled image summaries [17,47].

To address these challenges, recent work has increasingly explored vision-language pretraining using paired radiology reports as supervision. Contrastive approaches, including CLIP, CXR-CLIP, Language over Labels, and MedCLIP, learn transferable visual representations by aligning image and text embeddings [33,51,45,41]. These methods have demonstrated strong performance in medical imaging [4,17,47]. Global image-text alignment compresses report information into pooled representations, potentially limiting the supervision available for fine-grained findings. This limitation does not apply uniformly to contrastive learning: local and knowledge-guided approaches explicitly model finer-grained relationships. Our comparison focuses on the contrastive recipe evaluated here, rather than all possible contrastive formulations.

Autoregressive vision-language pretraining provides an alternative supervision paradigm. Rather than optimizing only global image-text agreement, autoregressive models condition on the image and generate textual outputs sequentially token by token [30,6,2,7,55,25]. This formulation naturally supports richer and more heterogeneous supervision from the same study, including full reports, impression summaries, structured findings, region descriptions, bounding-box annotations, question answering, and step-by-step clinical reasoning. Because supervision is provided through token-level prediction rather than a single pooled similarity objective, rare findings contribute explicit gradients whenever they appear, potentially encouraging representations that better preserve fine-grained anatomical and contextual information. Autoregressive training therefore offers a scalable framework for integrating diverse forms of clinical supervision within a unified objective, without requiring separate task-specific formulations. In ra-

diology, where interpretation depends not only on identifying abnormalities but also on understanding their anatomical location, prominence, and contextual relationships, multimodal next-token prediction may therefore produce more transferable visual representations for downstream long-tail recognition.

Despite growing interest in large autoregressive vision-language models, it remains unclear whether radiology-native autoregressive pretraining provides better downstream transfer than medical contrastive pretraining under controlled experimental conditions. Prior comparisons often differ simultaneously in architecture, pretraining corpus, optimization strategy, and downstream evaluation protocol, making it difficult to isolate the contribution of the pretraining objective itself. Furthermore, evaluations of medical vision-language models have primarily focused on aggregate discrimination performance, with comparatively less attention given to long-tail behavior and uncertainty-aware evaluation, both of which are critical for reliable clinical deployment [11,56,18,14].

In this work, we address this gap through a comparison of contrastive and autoregressive pretraining under a matched downstream architecture for long-tailed multi-label CXR classification. We compare MedCLIP-style contrastive pretraining and autoregressive vision-language pretraining using the InternViT architecture trained under each objective, followed by independent downstream fine-tuning with an ML-Decoder classification head, thereby controlling the downstream architecture while comparing complete pretraining recipes. Our autoregressive framework is trained in a radiology-native setting using heterogeneous supervision, including structured multi-section report templates, auxiliary abnormality-focused targets, and approximately 100k region annotations. To evaluate rare-finding transfer more rigorously, we additionally construct LLM-expanded label sets for MIMIC-CXR and CheXpert that increase the granularity of downstream evaluation. Experiments across an internal dataset, MIMIC-CXR, CheXpert, and PadChest demonstrate that autoregressive pretraining achieves stronger downstream discrimination on public benchmarks, improves performance on multiple tail abnormalities, and yields lower EAURC on the public datasets, while the contrastive baseline retains lower EAURC on the internal dataset [44,14].

Our contributions are threefold:

- We compare contrastive and autoregressive pretraining recipes for long-tailed chest X-ray classification, using separate InternViT encoders independently fine-tuned with the same ML-Decoder architecture.
- We develop a radiology-native autoregressive pretraining framework that integrates heterogeneous supervision from structured reports, auxiliary text targets, and region-level annotations, enabling the encoder to learn representations beyond global image-report alignment.
- We provide a comprehensive evaluation across internal and public CXR datasets, including head-tail analysis, LLM-expanded long-tail benchmarks for MIMIC-CXR and CheXpert, and uncertainty-aware assessment using risk-coverage metrics.

2 Related Work

2.1 Long-tailed chest X-ray classification

Long-tailed recognition in chest X-rays has received increasing attention, particularly following the introduction of the CXR-LT benchmark and challenge, which demonstrated that strong aggregate performance can conceal poor recognition of rare findings [15,16]. Existing approaches primarily address imbalance during supervised downstream training through techniques such as re-weighted losses, asymmetric objectives, resampling, data augmentation, and architectural modifications including multi-view fusion [54,36,31,24]. Although these methods improve classification under skewed label distributions, they focus largely on optimization at the downstream stage rather than on the representations learned during pretraining.

This limitation is particularly important in chest radiography, where rare abnormalities are often subtle, spatially localized, and difficult to distinguish using coarse global features alone. Consequently, there has been growing interest in leveraging radiology reports as an additional source of supervision, since textual descriptions provide clinically meaningful context that may help models learn more transferable representations for rare and fine-grained findings. This motivation has led to the increasing adoption of vision-language pretraining methods in medical imaging.

2.2 Contrastive pretraining

Medical vision-language pretraining commonly uses radiology reports as weak supervision for learning transferable visual representations. Early and widely adopted methods follow CLIP-style contrastive alignment, including CXR-CLIP, Language over Labels, and MedCLIP, which learn a shared image-text embedding space from paired or weakly paired image-report studies [33,51,45,41]. By incorporating report semantics during pretraining, these approaches improve representation quality beyond purely supervised image classification.

Building on this paradigm, subsequent work has explored richer forms of alignment to better capture the structure and granularity of radiology reports. For example, GLoRIA introduces both global and local alignment objectives between report words and image regions, while MedKLIP incorporates knowledge-enhanced triplet encoding derived from medical reports to guide visual learning from chest X-rays [17,4,47]. Similar trends have also emerged outside medical imaging, where methods such as DenseCLIP and FineCLIP improve dense and fine-grained recognition by modeling local visual-textual correspondences rather than relying solely on pooled embeddings [34,20]. In parallel, recent studies use large language models to extract structured disease information from reports, improving weak supervision and report understanding [43].

Collectively, these works suggest that richer supervision and finer-grained image-text interactions can improve representation learning. Global alignment remains a common component, although local and knowledge-guided contrastive

methods also exploit finer-grained supervision. This has motivated growing interest in autoregressive vision-language modeling, where supervision is provided through sequential next-token prediction rather than pooled embedding alignment alone.

2.3 Autoregressive pretraining

Autoregressive and generative objectives have emerged as a strong alternative for multimodal representation learning. Early work such as VirTex demonstrated that generating captions from semantically rich annotations can produce transferable visual features, while SimVLM introduced a simplified large-scale pre-training framework based on prefix language modeling [9,42]. Subsequent methods further expanded this paradigm. CoCa combines contrastive alignment with autoregressive caption generation within a unified framework, and Emu extends next-token prediction to interleaved image-text sequences [52,40]. More recently, AIM and AIMv2 showed that large autoregressively pretrained vision encoders can scale effectively and transfer well across classification, localization, and grounding tasks [10,12].

A key advantage of autoregressive supervision is that it naturally supports dense and heterogeneous training signals. Instead of optimizing a single global similarity objective, autoregressive models can learn from full reports, impression summaries, structured findings, region descriptions, and other clinically relevant textual outputs. This property is particularly appealing in radiology, where understanding subtle abnormalities often depends on contextual and localized information. In the medical domain, generative pretraining has therefore been explored for joint image-text understanding and report generation [30], while recent multimodal systems such as Florence-VL and InternVL3 further demonstrate the increasing role of autoregressive vision-language modeling [6,55].

Despite this progress, there remains limited evidence on whether autoregressive pretraining provides advantages over medical contrastive pretraining for long-tailed multi-label chest X-ray classification under controlled experimental settings. In particular, prior comparisons often differ simultaneously in architecture, optimization strategy, and downstream evaluation protocol, making it difficult to isolate the effect of the pretraining objective itself.

2.4 Uncertainty-aware evaluation in medical imaging

Beyond discrimination performance alone, reliable deployment of medical AI systems also requires robust uncertainty estimation. This is especially important in long-tailed classification settings, where rare findings are underrepresented during training and therefore more susceptible to unreliable predictions. Prior work in machine learning and radiology distinguishes between aleatoric uncertainty, which arises from inherent noise or ambiguity in the data, and epistemic uncertainty, which reflects uncertainty in the model parameters due to limited or biased observations [18,11]. In chest radiography, both forms of uncertainty are

clinically relevant because findings may be visually ambiguous, sparsely represented, or confounded by overlapping pathology.

Recent work has increasingly emphasized that uncertainty should be evaluated jointly with prediction reliability and clinical decision risk rather than treated as a secondary metric [56,38]. Beyond calibration alone, selective prediction and risk-coverage analysis assess whether a model’s confidence can appropriately rank its own errors, enabling uncertain predictions to be deferred to clinicians when necessary [14,13]. However, uncertainty-aware evaluation remains relatively underexplored in studies of medical vision-language pretraining, particularly in the long-tailed multi-label setting where rare diseases are both more difficult to recognize and more prone to overconfident failure modes.

3 Methodology

This study investigates how different vision-language pretraining paradigms influence representation learning and downstream transfer in long-tailed chest X-ray classification. The problem is particularly challenging because rare thoracic abnormalities are often subtle, sparsely represented, and frequently co-occur with more common findings [16,22]. In such settings, robust recognition depends on representations that preserve both localized visual detail and broader clinical context. The contrastive recipe considered here aligns pooled image and text representations, whereas autoregressive pretraining learns by sequentially predicting clinically grounded text conditioned on the image. This autoregressive formulation naturally supports heterogeneous supervision signals, including report generation, structured findings, region descriptions, question answering, and bounding-box annotations within a unified training framework. If such supervision improves representation quality, autoregressive pretraining may provide a scalable approach for incorporating increasingly diverse forms of clinical supervision. To evaluate these differences, we perform a comparison under matched downstream architecture between contrastive and autoregressive pretraining under matched downstream evaluation settings, described in the following sections.

3.1 Overview

We study the effect of the pretraining paradigm on long-tailed multi-label chest X-ray (CXR) classification by comparing two vision-language pretraining strategies under a controlled downstream setup: (i) MedCLIP-style contrastive pretraining [41], and (ii) autoregressive pretraining using an InternVL3-style vision-language model [55]. To control the downstream architecture, both approaches use independently pretrained instances of the same image-backbone family and are fine-tuned with the same classification head, namely ML-Decoder [35] shown in Figure 1.

Formally, let x denote a chest X-ray image and let $y \in \{0,1\}^C$ denote the multi-label target vector over C abnormalities. In both settings, a pretrained

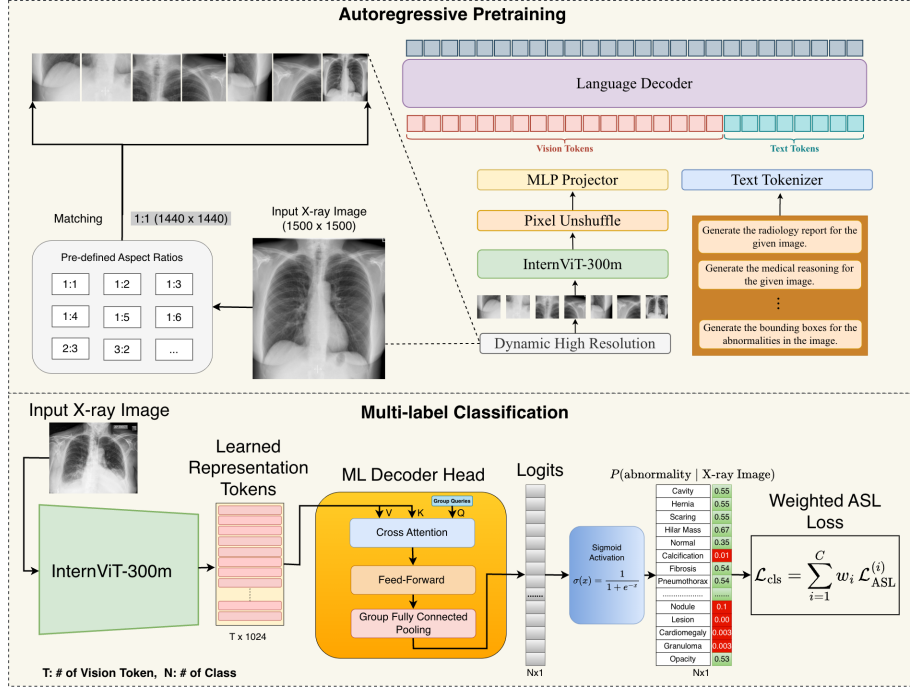


Fig. 1: Overview of the proposed framework. **Top:** In autoregressive pretraining, each chest X-ray is dynamically tiled at high resolution according to a matched aspect ratio and encoded by InternViT-300M. The visual features are transformed into vision tokens through pixel unshuffle and an MLP projector, then combined with prompt-conditioned text tokens in a language decoder. The model is trained with heterogeneous radiology-native supervision, including report generation, medical reasoning, structured abnormality outputs, and region-level descriptions such as abnormality bounding boxes. **Bottom:** For downstream long-tailed multi-label classification, the pretrained InternViT encoder is transferred and coupled with an ML-Decoder head, which predicts logits for all abnormalities and is optimized with weighted asymmetric loss.

vision encoder $E(\cdot)$ first maps the image to a set of visual features,

$$\mathbf{z} = E(x), \quad (1)$$

which are then passed to independently trained instances of the same multi-label classifier architecture $g(\cdot)$ to obtain per-label probabilities,

$$\hat{\mathbf{p}} = g(\mathbf{z}) \in [0, 1]^C. \quad (2)$$

The transferred encoders share an architecture but are pretrained with different objectives and supervision. In particular, the autoregressive recipe additionally uses structured and region-level targets. This design controls the downstream architecture but does not isolate the effect of the pretraining objective from the supervision, initialization, or optimization recipe.

3.2 Shared Vision Backbone Architecture

For both paradigms, we use InternViT-300M as the visual backbone. InternViT is a transformer-based image encoder designed for high-resolution visual understanding and defines the common encoder architecture in our comparison. In the autoregressive setup, the encoder operates within the InternVL3 framework [55]; in the contrastive setup, an independent encoder is trained using our MedCLIP-style contrastive recipe [41]. By transferring the same encoder family to the downstream task, we reduce architectural confounding while comparing the transfer performance of the two recipes.

3.3 Autoregressive Vision-Language Pretraining

Our autoregressive pretraining setup is implemented using InternVL3 [55], as illustrated at the top of Fig. 1, where an InternViT image encoder is trained jointly with a language decoder via conditional next-token prediction. Given an image x and a textual prompt q , the model generates a target sequence

$$t = (t_1, t_2, \dots, t_T), \quad (3)$$

which may correspond to a structured report, an impression-style summary, a compact abnormality list, or a region-grounded textual description. The training objective maximizes the conditional likelihood of the target sequence:

$$\mathcal{L}_{\text{AR}} = - \sum_{k=1}^T \log p(t_k \mid x, q, t_{<k}). \quad (4)$$

This objective couples visual representation learning with language generation. The image is encoded into a sequence of visual tokens that condition the decoder, and the encoder is trained to produce features that support accurate and coherent text generation. This motivates the hypothesis that the learned

representations retain structured and fine-grained information; classification results alone do not directly establish visual grounding.

A key advantage of this formulation is that it naturally supports multiple forms of supervision through prompt-conditioned generation. The same image can be used to generate different clinically meaningful outputs, such as detailed findings, concise impressions, abnormality summaries, or region-level descriptions. This provides dense, token-level supervision across multiple views of the data, rather than relying on a single global alignment signal.

To further enrich supervision, we construct diverse training targets from radiology reports and auxiliary annotations. These include: (i) section-wise report templates, (ii) short impression-like summaries, (iii) compact abnormality-focused outputs, and (iv) region-level descriptions derived from bounding-box annotations. Such heterogeneous supervision is intended to associate localized image patterns with corresponding text, potentially benefiting subtle and spatially specific abnormalities.

To study how the strength of the language modeling objective affects the learned visual features, we pretrain two models with identical InternViT encoders but different decoder scales. One model uses a 2B-parameter language decoder (Med-AR-2B), while the other uses an 8B-parameter decoder (Med-AR-8B). A larger decoder can model more complex linguistic patterns, but may also dominate training or overfit to dataset-specific text, potentially limiting the generality of the visual features. Whether decoder capacity changes the information retained by the vision encoder remains an empirical question.

After pretraining, the language decoder is discarded and only the vision encoder is transferred to the downstream classification task. Both Med-AR-2B and Med-AR-8B are evaluated under an identical fine-tuning setup and compared against an InternViT encoder pretrained with MedCLIP, enabling a comparison of the resulting encoders under matched downstream architecture.

3.4 MedCLIP-Style Contrastive Pretraining

Our contrastive baseline combines InternViT with BiomedVLP-CXR-BERT and uses the MedCLIP semantic-matching objective [41,3]. We retain the name Med-CLIP in the result tables for this implementation. Let $v = f_{\text{img}}(x)$ and $u = f_{\text{text}}(r)$ denote image and report embeddings. Medical semantic information from report-derived labels determines soft matching targets for image-text pairs, rather than assigning a positive target only to the paired report and treating every other report as a negative.

The bidirectional semantic-matching objective is

$$\mathcal{L}_{\text{MedCLIP}} = \frac{1}{2} (\mathcal{L}_{\text{img} \rightarrow \text{text}} + \mathcal{L}_{\text{text} \rightarrow \text{img}}), \quad (5)$$

where each directional term is a cross-entropy between the semantic matching targets and the similarity-based predicted matching distribution. This is distinct from standard paired InfoNCE: images and reports with related medical labels

can receive nonzero matching targets even when they are not an original pair. After pretraining, the visual encoder is transferred to downstream classification.

The baseline aligns pooled visual and textual representations with semantic supervision. The autoregressive recipe additionally uses structured generation targets and region annotations. The comparison therefore evaluates complete pretraining recipes rather than isolating the objective or establishing a limitation of all contrastive methods.

3.5 Downstream Fine-Tuning with ML-Decoder

After pretraining, all three vision encoders are fine-tuned on downstream multi-label classification using ML-Decoder [35] as shown in bottom of Fig 1. ML-Decoder is a transformer-based prediction head designed for efficient large-scale multi-label recognition. Instead of learning an independent classifier for each label, it uses a fixed set of learnable queries that interact with the encoded image features to produce label predictions.

Let $\mathbf{z} \in \mathbb{R}^{N \times d}$ denote the output token sequence of the vision encoder, where N is the number of visual tokens and d is the feature dimension. ML-Decoder maps $\mathbf{z}$ to class logits

$$\mathbf{s} = h(\mathbf{z}) \in \mathbb{R}^C, \quad (6)$$

followed by sigmoid activation to obtain class probabilities

$$\hat{p}_i = \sigma(s_i), \quad i = 1, \dots, C. \quad (7)$$

A key advantage of ML-Decoder is that it scales efficiently to large label spaces while still modeling label dependencies and co-occurrence structure. This makes it particularly suitable for long-tailed CXR classification, where abnormalities often co-occur and the number of classes can be large in fine-grained evaluation settings.

3.6 Loss Function for Long-Tailed Multi-Label Classification

To address both inter-class and intra-class imbalance during downstream training, we combine class-specific weighting with asymmetric loss (ASL) [36,24].

Class-specific weighting. Let $y_i \in \{0, 1\}$ denote the ground-truth label for class i , and let ρ_i denote the positive sample ratio of class i in the training set. We define a class-aware sample weight as

$$w_i = y_i \exp(1 - \rho_i) + (1 - y_i) \exp(\rho_i). \quad (8)$$

This weighting increases the contribution of rare positive classes while preserving the contribution of negative examples.

Asymmetric loss. For each class i , the asymmetric loss is defined as

$$\mathcal{L}_{\text{ASL}}^{(i)} = -y_i(1 - \hat{p}_i)^{\gamma_+} \log(\hat{p}_i) - (1 - y_i) \hat{p}_{m,i}^{\gamma_-} \log(1 - \hat{p}_{m,i}), \quad (9)$$

where

$$\hat{p}_{m,i} = \max(\hat{p}_i - m, 0). \quad (10)$$

Here, γ_+ and γ_- are focusing parameters for positive and negative samples, respectively, and m is a probability margin used to suppress easy negatives. Following prior long-tailed CXR classification work [24], we use $\gamma_+ = 1$, $\gamma_- = 4$, and $m = 0.05$.

Final training objective. The final downstream loss is computed as the weighted sum over classes:

$$\mathcal{L}_{\text{cls}} = \sum_{i=1}^C w_i \mathcal{L}_{\text{ASL}}^{(i)} \quad (11)$$

This objective jointly addresses inter-class imbalance, by up-weighting rare classes, and intra-class imbalance, by down-weighting easy negatives. As a result, it is well suited for long-tailed multi-label CXR classification.

3.7 Long-tailed evaluation protocol

We study the downstream task as multi-label abnormality classification on four datasets: an internal clinical dataset, MIMIC-CXR, CheXpert and PadChest. This multi-dataset setup is important because long-tailed behavior can be sensitive to dataset composition, reporting style, disease prevalence, and annotation granularity. Evaluating on both internal and public datasets therefore helps distinguish in-domain gains from broader transfer behavior.

To make the long-tail analysis explicit, we partition labels separately within each dataset according to their frequency in the *training split*. Let c_i denote the number of positive training examples for label i . Because the label-frequency distribution is strongly long-tailed, we first transform counts into log space:

$$x_i = \log(1 + c_i). \quad (12)$$

We then compute dataset-specific thresholds from the empirical 33rd and 67th percentiles of the log-count distribution:

$$T_1 = Q_{0.33}(x), \quad T_2 = Q_{0.67}(x). \quad (13)$$

To obtain interpretable thresholds in the original count space, we map these quantiles back using the inverse transform:

$$C_1 = \exp(T_1) - 1, \quad C_2 = \exp(T_2) - 1. \quad (14)$$

Finding labels are then assigned to three groups directly according to their positive training counts:

$$\text{tail} : c_i < C_1, \quad \text{medium} : C_1 \leq c_i < C_2, \quad \text{head} : c_i \geq C_2. \quad (15)$$

This procedure provides a dataset-adaptive partition of the label space while applying the same rule across all benchmarks. Because the log transform preserves ordering, the groups principally reflect relative label-frequency ranks within each dataset, with boundary details determined by quantile interpolation and ties. Tail membership therefore does not imply the same absolute prevalence across datasets. The resulting head/medium/tail split enables finer-grained analysis of whether improvements from pretraining are concentrated on frequent labels or extend to intermediate- and low-frequency abnormalities. The dataset-specific count thresholds used in this work are summarized in Table 2.

For MIMIC-CXR and CheXpert, we further expand the evaluation space through an LLM-based label extraction pipeline, producing a finer-grained long-tailed benchmark than the standard label set. This expanded label space, shown in Figure 2 and Figure 3, is not introduced as a separate task; rather, it serves as a more sensitive probe of representation quality under rare-label transfer, where differences between pretraining paradigms may be difficult to detect using only a small number of coarse disease categories.

4 Experiments

4.1 Dataset Preparation

Pre-training: For pre-training we use an internal training dataset of 1.4 million chest X-rays and their corresponding reports, obtained from various hospitals, anonymized and cleaned. For MedCLIP pretraining, we leverage paired image-text data and use prompt engineering to design an effective instruction for extracting detailed disease labels from radiology reports, enabling fine-grained annotation of more than 150 findings, including rare abnormalities. The LLM (Qwen3 [49]) was prompted to produce a JSON object listing all present findings, guided by a curated list of over 150 finding categories. Downstream label expansion is described separately in Section 4.2. For autoregressive pre-training, we create a structured JSON template consisting of 15 subheadings covering the various anatomical regions to look at in a chest radiograph. Under each subheading, we include further categorizations for abnormalities. In addition to the 15 headings, we also add *Findings*, *Impression*, *Summary*, *Reasoning* and *Prominence Score* sections. The prominence score section provides a dictionary of the report-derived findings along with a generated prominence score, from 0–5. We use Qwen3 to extract this information from the reports and output the response in the structured template format. In cases when an abnormality is not mentioned or is not present, we instruct the LLM to use a negation statement. This procedure conflates non-mention with explicit absence and is a source of weak-supervision noise. The reasoning and prominence fields are generated from reports; they are not independently adjudicated visual assessments. Furthermore, we incorporate expert-annotated bounding boxes for approximately 20 thoracic findings (e.g., opacity, nodule, pneumothorax, mass), represented as four-point coordinates. For each supervision type: global report parsing, structured abnormality extraction, prominence scoring, and region-level bounding-box, we design

dedicated prompt templates for Qwen3, ensuring consistent JSON outputs across these heterogeneous data sources which are then used for pretraining InternVL3 model.

Finetuning: We conduct separate training experiments on four large-scale CXR datasets: an internal collection of chest radiographs (Internal), PadChest [5], CheXpert [19] and MIMIC-CXR [21]; the splits in Table 1 contain 1,872,398 training images and 297,608 test images in total. For each dataset, models are trained exclusively on its corresponding training split and evaluated on its own held-out test set, as summarized in Table 1.

- **Internal Dataset:** The training split includes approximately 1.4M chest X-rays from multiple hospitals. An LLM fine-tuned on radiologist annotations extracts 40 report-derived labels, including hilar mass.
- **PadChest:** Includes 160K CXRs labeled with 174 UMLS-derived findings. We use the 29 most frequent labels for training and the held-out split reported in Table 1. Volume loss is the least prevalent retained finding. Tail analysis is restricted to these retained labels, rather than the full PadChest vocabulary.
- **MIMIC-CXR:** A public dataset with 173K CXRs (frontal) and reports. Our LLM extracts ~ 150 tags for granularity. We use 110 finding labels (with training prevalence > 0.0002) for training. Bullae is the least common retained finding. We plan to release the derived labels and split identifiers subject to the source dataset terms.
- **CheXpert:** A public dataset with approximately 160K retained CXRs (frontal) and reports. Our LLM extracts ~ 150 tags for granularity. We use 115 finding labels (with training prevalence > 0.0002) for training. Left ventricular assist device (LVAD) is the least common retained label. We plan to release the derived labels and split identifiers subject to the source dataset terms.

Dataset	Training Samples	Testing Samples
Internal	1,432,373	242,965
MIMIC-CXR	152,240	21,466
PadChest	138,235	22,626
CheXpert	149,550	10,551

Table 1: Dataset statistics, showing the number of training and testing samples for each dataset used in fine-tuning.

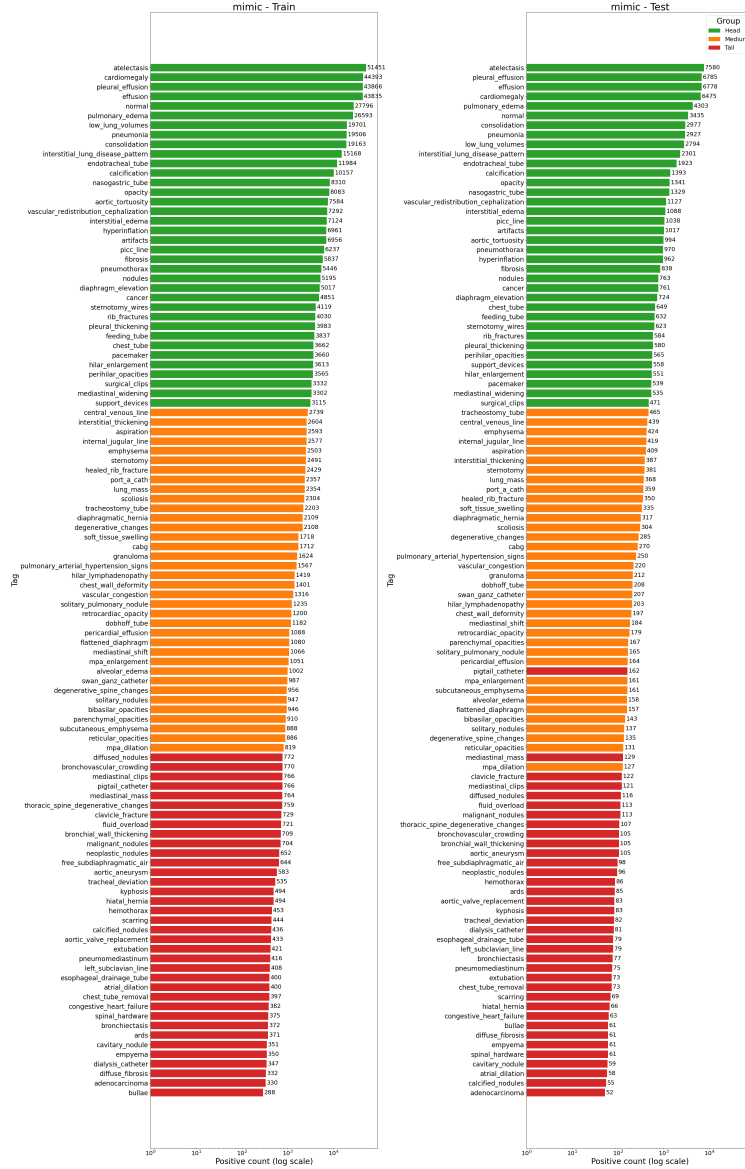


Fig. 2: Distribution of extracted MIMIC-CXR tags. The classification label space contains 110 labels; available per-label results are listed in Supplementary Section S5. The data is split into head, medium, and tail tags. Train prevalence $\geq 0.02\%$; log-count thresholds: tail < 1106 , medium < 4210 , head ≥ 4210 .

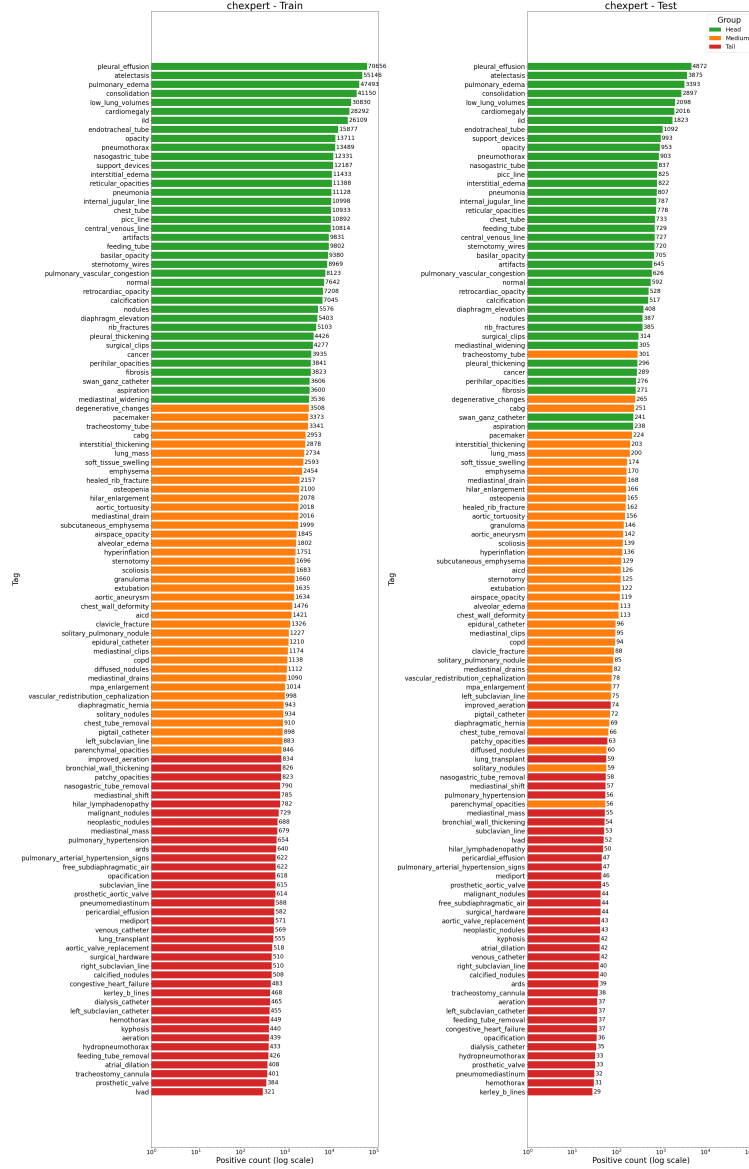
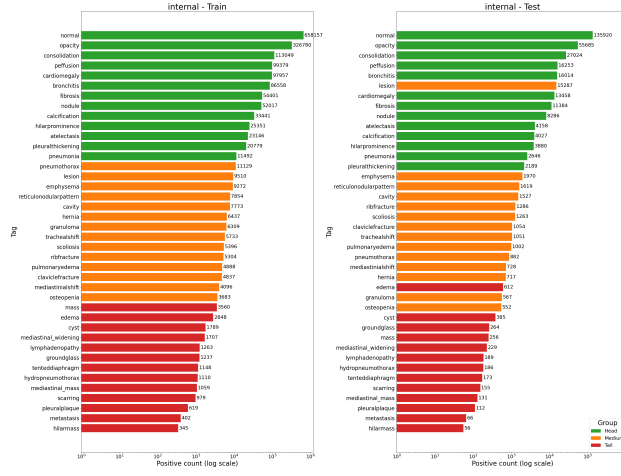
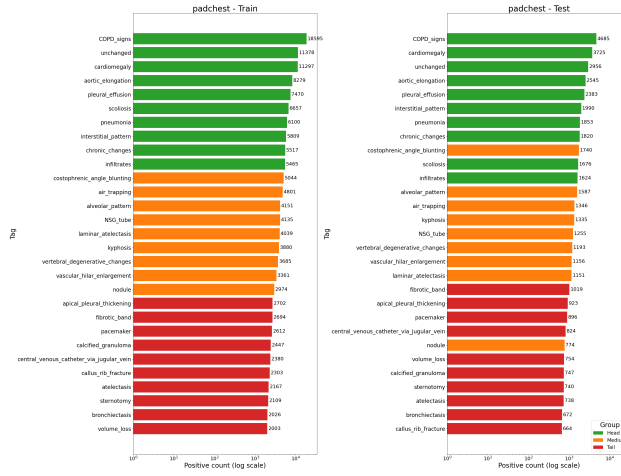


Fig. 3: Distribution of the CheXpert tag extraction dataset with 115 tags. The data is split into head, medium, and tail tags. Train prevalence $\geq 0.02\%$; log-count thresholds: tail < 841 , medium < 3519 , head ≥ 3519 .



(a) Distribution of the Internal dataset with 40 tags. The data is split into head, medium, and tail tags. Train prevalence $\geq 0.02\%$; log-count thresholds: tail < 3667 , medium < 11176 , head ≥ 11176 .



(b) Distribution of the PadChest dataset with 29 tags. The data is split into head, medium, and tail tags. Train prevalence $\geq 0.02\%$; log-count thresholds: tail < 2765 , medium < 5361 , head ≥ 5361 .

Fig. 4: Tag distribution comparison for the Internal and PadChest datasets.

The distribution of each tag in the train and test set is given in Figures 2 3 4a and 4b and training set thresholds for grouping into head, medium and tail for each dataset is mentioned in Table 2 and discussed in Sec 3.7.

Table 2: Dataset-specific thresholds used to partition labels into head, medium, and tail groups based on the number of positive training examples. Thresholds are derived from the long-tailed label distribution of each dataset after filtering labels with training prevalence $\geq 0.02\%$ as discussed in Section 3.7.

Dataset	Tail	Medium	Head
Internal	$c_i < 3667$	$3667 \leq c_i < 11176$	$c_i \geq 11176$
MIMIC-CXR	$c_i < 1106$	$1106 \leq c_i < 4210$	$c_i \geq 4210$
PadChest	$c_i < 2765$	$2765 \leq c_i < 5361$	$c_i \geq 5361$
CheXpert	$c_i < 841$	$841 \leq c_i < 3519$	$c_i \geq 3519$

4.2 LLM-based Label Extraction and Concordance Analysis

To expand the label space of MIMIC-CXR and CheXpert, we derived structured labels from free-text radiology reports using a large language model (LLM). The prompt specified a comprehensive set of thoracic abnormalities and required the model to return their presence or absence in a structured JSON format. Positive findings were restricted to those supported by the report, with explicit instructions to avoid inference beyond the documented text. We additionally included an `other_findings` field to capture abnormalities outside the predefined vocabulary. While this improved coverage, it introduced lexical variability across models, including alternative terminology and singular/plural forms; these free-form outputs were therefore mapped to a canonical label vocabulary during post-processing.

We evaluated inter-model extraction concordance on a 1,000-report subset using seven LLMs: DeepSeek V4 Pro, Gemini 3.1 Pro, Gemini 3 Flash, GLM 5.1, Qwen3, Claude Haiku, and GPT-5 Mini. Because the resulting multi-label matrix was highly sparse, with approximately 98% of report-label pairs corresponding to shared negatives, raw percentage agreement would substantially overestimate concordance. We therefore evaluated chance-corrected agreement using Cohen’s κ for pairwise model comparisons and Fleiss’ κ for agreement across the complete seven-model panel. Mean pairwise Cohen’s κ was approximately 0.83 at the individual-label level and increased to 0.87 after related labels were collapsed into parent finding groups, indicating that a substantial component of apparent disagreement resulted from differences in label granularity. Consistently, parent-group Fleiss’ κ showed strong seven-model concordance across the clinical finding groups ($\kappa = 0.79\text{--}0.98$), whereas `other_findings` showed lower agreement ($\kappa = 0.56$), reflecting its more heterogeneous and open-ended vocabulary (Figure 5).

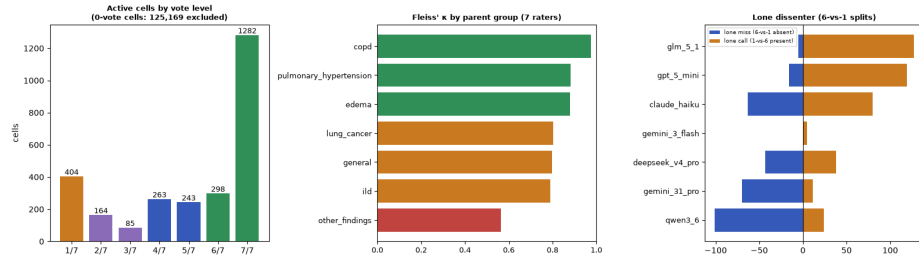


Fig. 5: Inter-model concordance of LLM-derived labels. **Left:** Distribution of votes across seven models for active report–tag pairs, where the vote denotes the number of models (1–7) identifying a finding as present; unanimous-negative pairs (0/7) are excluded. **Middle:** Fleiss’ κ by parent finding group, showing strong agreement across clinical categories. **Right:** Model-specific 6-vs-1 disagreements, distinguishing lone misses from lone positive calls.

Residual disagreement was dominated by *synonym routing*, whereby models identified the same underlying finding but assigned different keys. Common examples included `effusion` versus `pleural_effusion` and the nodule family (`nodules`, `solitary_nodules`, and `solitary_pulmonary_nodule`). When `solitary_pulmonary_nodule` was missed under its specific key, `nodules` was assigned in 96% of such cases; similarly, `pleural_effusion` was assigned in 95% of cases in which `effusion` was missed. The open-ended `other_findings` field further contributed to this naming variability, reinforcing the need for canonicalization. One notable exception was `artifacts`, which exhibited model-specific over-extraction rather than synonym substitution and consequently behaved as a source of annotation noise.

For large-scale downstream label expansion, we selected Gemini 3 Flash for its inclusive extraction behavior relative to the other models. This stage is separate from Qwen3-based construction of the internal pretraining targets. General and specific labels may both be emitted and are examined during canonicalization. Without expert reference labels, this comparison does not establish clinical precision or recall. Inter-model agreement measures reproducibility of report extraction, and shared model errors can remain undetected. Parent-group agreement may also conceal errors in fine-grained distinctions. The expanded labels should therefore be interpreted as report-derived weak supervision rather than independently validated image-level ground truth. Supplementary Section S1.2 details parsing, normalization, and concordance analysis.

4.3 Implementation

Pretraining MedCLIP is pretrained using AdamW (lr=2e-5) with a one-cycle scheduler. The text encoder is BiomedVLP-CXR-BERT, and the image encoder is InternViT (embedding size=512).

For training the VLM, we use a learning rate of 1e-5 with AdamW and a warmup

ratio of 0.05.

Fine-Tuning ML-Decoder with $k=N$ (no. of classes), where k is size of group queries serves as the classification head, processing the final hidden state of the image encoder (embedding size = 1024). The model is fine-tuned using Asymmetric Loss, AdamW, and a OneCycle scheduler (batch size = 2304, max lr = $5e-05$). We used 8XH100 GPUs for both MedCLIP and Med-AR training.

4.4 Evaluation Metrics and Statistical Analysis

Evaluation metrics. We evaluate downstream performance using AUROC, AUPRC, sensitivity, specificity, and Excess Area Under the Risk–Coverage Curve (EAURC). Since the task is multi-label classification, each abnormality is treated as a one-vs-rest binary prediction problem, and metrics are computed per label before being averaged over all labels as well as over the head, medium, and tail subsets.

We report AUROC as a threshold-independent measure of ranking quality, and AUPRC as a complementary metric that is especially informative under class imbalance, where positive cases may be rare [37]. To characterize thresholded operating performance, we also report sensitivity and specificity, which quantify the ability to recover positive cases while avoiding false alarms. For public-dataset operating-point results, thresholds are selected per label by maximizing Youden’s J on the validation set, as described in Supplementary Section S2.1.

To assess uncertainty-aware reliability, we use EAURC, a selective-prediction metric that evaluates how well model confidence ranks predictions by difficulty [14]. Lower EAURC indicates a lower excess selective risk under the evaluated confidence ordering. This metric does not measure probability calibration or establish clinical deployment readiness; in sparse multi-label tasks, ordinary error risk can be dominated by negative examples.

Statistical significance tests. We use two complementary statistical tests to assess whether differences between pretraining paradigms are systematic. First, we apply the Wilcoxon signed-rank test to paired per-label AUROC and AUPRC values [46]. This test is appropriate because the comparison is paired at the label level, and the distribution of per-label metric differences is not guaranteed to be Gaussian. It therefore provides a robust non-parametric assessment of whether one pretraining strategy tends to outperform the other across all labels, or within the head and tail subsets.

Second, we use DeLong’s test to compare label-wise AUROCs between models [8]. DeLong’s test is designed for correlated ROC curves and is well suited here because the competing models are evaluated on the same test cases for each label. Unlike the Wilcoxon test, which summarizes paired differences across labels, DeLong’s test identifies whether the AUROC gap for a specific abnormality is statistically significant.

The reported significance comparisons are interpreted as nominal, exploratory evidence: correlated labels, multiple comparisons, repeated studies, and training variability limit inference. Non-significance does not establish

equivalence, and smaller p -values do not measure larger effect sizes. Together, these tests provide complementary evidence: the Wilcoxon signed-rank test assesses whether one pretraining paradigm is consistently better across labels, while DeLong’s test identifies which individual abnormalities show significant AUROC differences.

5 Results and Discussion

5.1 Main quantitative comparison

Table 3 summarizes mean AUROC and mean AUPRC across overall, head, medium, and tail label groups for the four evaluation benchmarks. A clear pattern emerges on the three public datasets, although the preferred autoregressive variant differs across them.

On PadChest, autoregressive pretraining consistently outperforms Med-CLIP across all prevalence groups, with Med-AR-2B achieving the best results throughout. Specifically, mean AUROC improves from 0.8657 to 0.8800 overall, from 0.8425 to 0.8571 for head labels, from 0.8531 to 0.8674 for medium labels, and from 0.9002 to 0.9143 for tail labels. The corresponding AUPRC values also increase consistently, from 0.4310 to 0.4595 overall, from 0.4428 to 0.4706 for head labels, from 0.3678 to 0.3941 for medium labels, and from 0.4762 to 0.5072 for tail labels.

A similar but even stronger trend is observed on MIMIC-CXR, where Med-AR-8B is the best-performing model across all groups. Overall mean AUROC increases from 0.8777 for Med-CLIP to 0.8971 for Med-AR-8B, while overall mean AUPRC rises from 0.2315 to 0.2769. The gains are visible across all prevalence regimes: for head labels, AUROC improves from 0.8756 to 0.8918 and AUPRC from 0.4010 to 0.4432; for medium labels, AUROC improves from 0.8841 to 0.9028 and AUPRC from 0.1902 to 0.2435; and for tail labels, AUROC improves from 0.8733 to 0.8967 and AUPRC from 0.1033 to 0.1441.

CheXpert provides a complementary test of this trend and highlights a stronger role for decoder scale. Med-AR-8B achieves the best AUROC and AUPRC in every prevalence group, lifting overall AUROC from 0.8585 (Med-CLIP) to 0.8759 and overall AUPRC from 0.2142 to 0.2391, with consistent gains for head labels (0.8433 to 0.8565 AUROC; 0.3710 to 0.3978 AUPRC), medium labels (0.8790 to 0.8920 AUROC; 0.1802 to 0.2054 AUPRC), and tail labels (0.8526 to 0.8788 AUROC; 0.0922 to 0.1151 AUPRC). In contrast, Med-AR-2B does not surpass Med-CLIP on this benchmark and is lowest across every prevalence group for both AUROC and AUPRC. Autoregressive supervision therefore remains beneficial on CheXpert, but the advantage is concentrated in the larger decoder, and a smaller autoregressive model is insufficient to outperform a strong contrastive baseline on this dataset.

As shown in Section 5.2, paired per-label Wilcoxon results support most of the Med-AR-8B improvements on the public datasets at the nominal 0.05 level. The exception is PadChest medium-label AUPRC ($p = 0.0645$). These tests

Table 3: Mean AUROC and AUPRC for overall, head, medium, and tail tags across datasets. The best value within each metric block is shown in bold, and the second-best value is underlined.

Dataset	Group	mAUROC			mAUPRC		
		Med-CLIP	Med-AR-8B (ours)	Med-AR-2B (ours)	Med-CLIP	Med-AR-8B (ours)	Med-AR-2B (ours)
Internal	Overall	0.9081	<u>0.9111</u>	0.9135	0.2862	0.2754	<u>0.2842</u>
	Head	0.8947	<u>0.8983</u>	0.9008	0.4289	<u>0.4367</u>	0.4415
	Medium	0.9069	<u>0.9100</u>	0.9124	<u>0.2434</u>	0.2292	0.2452
	Tail	0.9227	<u>0.9253</u>	0.9274	0.1897	0.1639	<u>0.1689</u>
PadChest	Overall	0.8657	<u>0.8722</u>	0.8800	0.4310	<u>0.4469</u>	0.4595
	Head	0.8425	<u>0.8497</u>	0.8571	0.4428	<u>0.4583</u>	0.4706
	Medium	0.8531	<u>0.8595</u>	0.8674	0.3678	<u>0.3818</u>	0.3941
	Tail	0.9002	<u>0.9060</u>	0.9143	0.4762	<u>0.4940</u>	0.5072
MIMIC	Overall	0.8777	0.8971	<u>0.8859</u>	0.2315	0.2769	<u>0.2541</u>
	Head	0.8756	0.8918	<u>0.8846</u>	0.4010	0.4432	<u>0.4264</u>
	Medium	0.8841	0.9028	<u>0.8881</u>	0.1902	0.2435	<u>0.2175</u>
	Tail	0.8733	0.8967	<u>0.8851</u>	0.1033	0.1441	<u>0.1185</u>
CheXpert	Overall	<u>0.8585</u>	0.8759	0.8422	<u>0.2142</u>	0.2391	0.2039
	Head	<u>0.8433</u>	0.8565	0.8310	<u>0.3710</u>	0.3978	0.3614
	Medium	<u>0.8790</u>	0.8920	0.8586	<u>0.1802</u>	0.2054	0.1682
	Tail	<u>0.8526</u>	0.8788	0.8365	<u>0.0922</u>	0.1151	0.0829

describe the observed label-wise differences and do not isolate the contribution of the pretraining objective.

The internal benchmark presents a more nuanced picture and is best described as broadly comparable between the two paradigms, with a modest edge for autoregressive pretraining in AUROC but a mixed pattern in AUPRC. In terms of AUROC, Med-AR-2B performs best across all groups, improving from 0.9081 to 0.9135 overall, from 0.8947 to 0.9008 for head labels, from 0.9069 to 0.9124 for medium labels, and from 0.9227 to 0.9274 for tail labels. However, the AUPRC results are less uniform. Med-AR-2B achieves the best performance for the head and medium groups, increasing AUPRC from 0.4289 to 0.4415 and from 0.2434 to 0.2452, respectively, whereas Med-CLIP remains strongest overall and on the tail subset, with AUPRC values of 0.2862 overall and 0.1897 on tail labels. Consistent with this mixed aggregate pattern, the statistical analysis in subsection 5.2 shows that several internal comparisons do not reach significance, especially for AR-8B and for the tail subset, supporting the interpretation that Med-AR and Med-CLIP are broadly comparable on the internal benchmark rather than cleanly separated.

The preferred autoregressive variant is dataset-dependent. Med-AR-2B achieves the highest AUROC on Internal and the best discrimination on PadChest, whereas Med-AR-8B is strongest on MIMIC-CXR and CheXpert. These results indicate that decoder scale interacts with the training recipe and downstream dataset; they do not establish the objective itself as the primary source of improvement.

5.2 Statistical significance across labels

To determine whether the aggregate improvements in Table 3 are systematic across labels rather than driven by a small subset of abnormalities, we applied the Wilcoxon signed-rank test to paired per-label AUROC and AUPRC values. Table 4 reports results for the overall, head, medium, and tail groups, comparing both Med-AR-8B and Med-AR-2B against Med-CLIP.

The public benchmarks generally support the performance trends in Table 3. On PadChest, both Med-AR variants have nominally significant overall gains in AUROC and AUPRC. Med-AR-2B also has nominally significant gains in every subgroup; Med-AR-8B does so except for medium-label AUPRC ($p = 0.0645$). On MIMIC-CXR, both variants have nominally significant AUROC and AUPRC gains in all groups. Effect magnitudes should be read from the metric differences, rather than inferred from the relative sizes of the p -values.

CheXpert shows the most asymmetric pattern between the two AR variants. Med-AR-8B achieves highly significant improvements over Med-CLIP across the overall benchmark and all prevalence groups, with AUROC p -values ranging from $2.17\text{e-}15$ overall to $3.01\text{e-}06$ on the tail subset, and AUPRC p -values from $1.08\text{e-}12$ overall to $1.27\text{e-}03$ on the tail subset. In contrast, Med-AR-2B fails to significantly outperform Med-CLIP on CheXpert: p -values are near 1.00 for all overall, head, and medium comparisons, and no significant AUPRC gain is observed on any subset. This directly corroborates the aggregate finding that Med-AR-2B does not surpass Med-CLIP on CheXpert and shows that the behavior is systematic at the per-label level rather than a metric-averaging artifact. Taken together with PadChest and MIMIC-CXR, these results indicate that the overall advantage of autoregressive pretraining on public benchmarks is broad and label-consistent, but depends critically on decoder scale: Med-AR-8B has nominally significant overall gains on all three public datasets, whereas Med-AR-2B has such gains on PadChest and MIMIC-CXR but not CheXpert.

The internal dataset presents a more nuanced picture, consistent with the broadly comparable aggregate results reported in the previous subsection. For Med-AR-8B, none of the Wilcoxon comparisons reach significance, indicating that its small aggregate improvements over Med-CLIP are not consistently expressed at the per-label level. In contrast, Med-AR-2B shows statistically significant AUROC gains for the overall, head, and medium groups, and also achieves a significant AUPRC improvement on the head subset. However, neither autoregressive variant shows significant gains on the internal tail subset, and most AUPRC comparisons remain non-significant. This pattern supports a balanced interpretation of the internal benchmark: Med-AR remains competitive and often favorable, particularly for AUROC and for the head and medium groups, but the differences relative to Med-CLIP are not uniformly strong enough to support the same clear superiority observed on the public datasets.

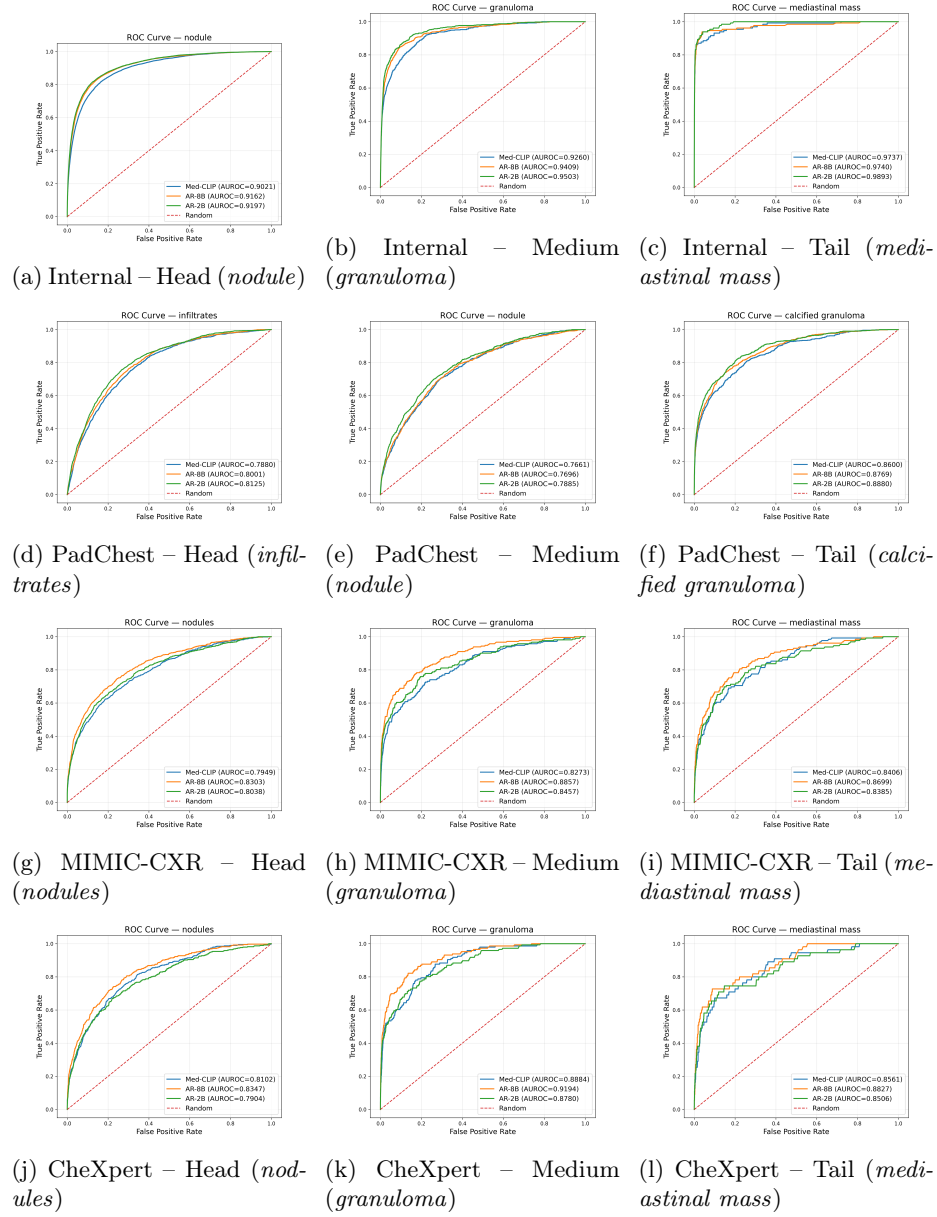


Fig. 6: ROC curves comparing Med-CLIP, Med-AR-8B, and Med-AR-2B across four datasets. Rows correspond to Internal, PadChest, MIMIC-CXR, and CheXpert, while columns correspond to representative head-, medium-, and tail-prevalence abnormalities, respectively.

Table 4: Wilcoxon signed-rank test results comparing autoregressive and CLIP-based pretraining using per-label AUROC and AUPRC across datasets for all, head, medium, and tail tags.

Dataset	Tag Group	Model	AUROC Stat.	AUROC p-val	AUPRC Stat.	AUPRC p-val
Internal	All	AR-8B	352.0	0.1578	533.0	0.9080
		AR-2B	215.0	0.0023	495.0	0.7982
	Head	AR-8B	27.0	0.0594	39.0	0.2131
		AR-2B	17.0	0.0123	25.0	0.0453
	Medium	AR-8B	31.0	0.1698	64.0	0.9045
		AR-2B	21.0	0.0471	57.0	0.7928
	Tail	AR-8B	57.0	0.6196	82.0	0.9710
		AR-2B	37.0	0.1788	90.0	0.9933
PadChest	All	AR-8B	48.0	4.36e-05	43.0	2.35e-05
		AR-2B	1.0	3.73e-09	4.0	1.30e-08
	Head	AR-8B	6.0	0.0137	9.0	0.0322
		AR-2B	0.0	9.77e-04	0.0	9.77e-04
	Medium	AR-8B	3.0	0.0098	9.0	0.0645
		AR-2B	0.0	0.00195	0.0	0.00195
	Tail	AR-8B	7.0	0.0186	1.0	0.00195
		AR-2B	1.0	0.00195	1.0	0.00195
MIMIC	All	AR-8B	206.0	1.04e-17	172.0	4.33e-18
		AR-2B	1032.0	8.42e-10	1164.0	8.90e-09
	Head	AR-8B	0.0	1.46e-11	1.0	2.91e-11
		AR-2B	13.0	1.28e-09	60.0	1.43e-06
	Medium	AR-8B	10.0	1.56e-10	16.0	6.15e-10
		AR-2B	189.0	0.0038	158.0	7.75e-04
	Tail	AR-8B	53.0	6.49e-07	43.0	1.88e-07
		AR-2B	127.0	4.08e-04	185.0	0.00959
CheXpert	All	AR-8B	524.0	2.17e-15	818.0	1.08e-12
		AR-2B	5418.0	1.00e+00	4175.0	0.9905
	Head	AR-8B	3.0	1.82e-11	31.0	8.64e-09
		AR-2B	643.0	1.00e+00	555.0	0.9969
	Medium	AR-8B	72.0	6.41e-07	92.0	4.08e-06
		AR-2B	671.0	1.00e+00	477.0	0.8876
	Tail	AR-8B	81.0	3.01e-06	167.0	0.00127
		AR-2B	549.0	0.9958	408.0	0.7072

Overall, the Wilcoxon analysis reinforces the central narrative of the paper. On PadChest, MIMIC-CXR, and CheXpert, autoregressive pretraining—most consistently through Med-AR-8B—has nominally significant improvements over Med-CLIP in most prevalence-group comparisons, whereas on the internal benchmark the two paradigms are better characterized as broadly comparable, with the strongest support for Med-AR concentrated in Med-AR-2B and in the higher-prevalence groups.

5.3 Label-wise AUROC differences

We further examined class-wise behavior using DeLong’s test for paired AUROCs. Figures 7 and 8 visualize the AUROC difference between autoregressive pretraining and Med-CLIP for each abnormality, separately for the head,

medium, and tail subsets in the Internal, MIMIC-CXR, PadChest, and CheXpert benchmarks. Statistically significant differences under DeLong’s test are marked in the figures.

The public datasets again show a clear advantage for autoregressive pre-training. On MIMIC-CXR, the majority of labels across the head, medium, and tail groups show positive AUROC differences, and many of these gains are statistically significant. The largest improvements are particularly evident among several medium and tail abnormalities, supporting the view that autoregressive supervision is especially beneficial when clinically relevant findings are less frequent or semantically more specific. PadChest exhibits a similarly favorable pattern, with consistently positive shifts across all three prevalence groups and multiple significant improvements among both common and infrequent abnormalities. CheXpert follows the same trend when compared against Med-AR-8B: positive AUROC differences dominate the head, medium, and tail subsets, and a substantial fraction of labels reach statistical significance, including several rare abnormalities. This label-level evidence complements the Wilcoxon analysis and confirms that the Med-AR-8B gains on CheXpert are distributed across many abnormalities rather than concentrated in a handful of high-prevalence classes. Consistent with the earlier aggregate and Wilcoxon findings, Med-AR-2B on CheXpert does not produce a comparable distribution of positive significant differences, again indicating that the CheXpert benefit is tied to decoder scale.

The internal dataset is more heterogeneous. In the head and medium groups, the class-wise AUROC differences mostly favor autoregressive pretraining, consistent with the Wilcoxon results for AR-2B. However, the internal tail subset remains mixed, with some abnormalities benefiting from AR and others continuing to favor Med-CLIP. This label-wise heterogeneity explains why the internal benchmark, despite favorable aggregate metrics for AR, yields more limited statistical support at the tail level. Overall, the DeLong analysis reinforces the central pattern of the paper: autoregressive pretraining produces broad and consistent gains on the public datasets, while the internal benchmark shows a more selective advantage concentrated in higher-prevalence groups and clinically important individual findings.

5.4 Uncertainty-aware evaluation

Table 5 reports Excess Area Under the Risk–Coverage Curve (EAURC), where lower values indicate better uncertainty ranking and, therefore, more effective selective prediction. This analysis complements the discrimination results by testing whether model confidence is informative about prediction correctness, rather than only whether the predicted scores separate positives from negatives.

The public datasets again favor autoregressive pretraining, although the best-performing variant varies across them. On MIMIC-CXR, Med-AR-8B achieves the lowest EAURC across all groups, reducing the overall value from 0.196 for Med-CLIP to 0.169, with consistent improvements for head labels (0.365 to 0.334), medium labels (0.186 to 0.146), and tail labels (0.038 to 0.027). Med-AR-2B also improves over Med-CLIP in every group, although it remains slightly

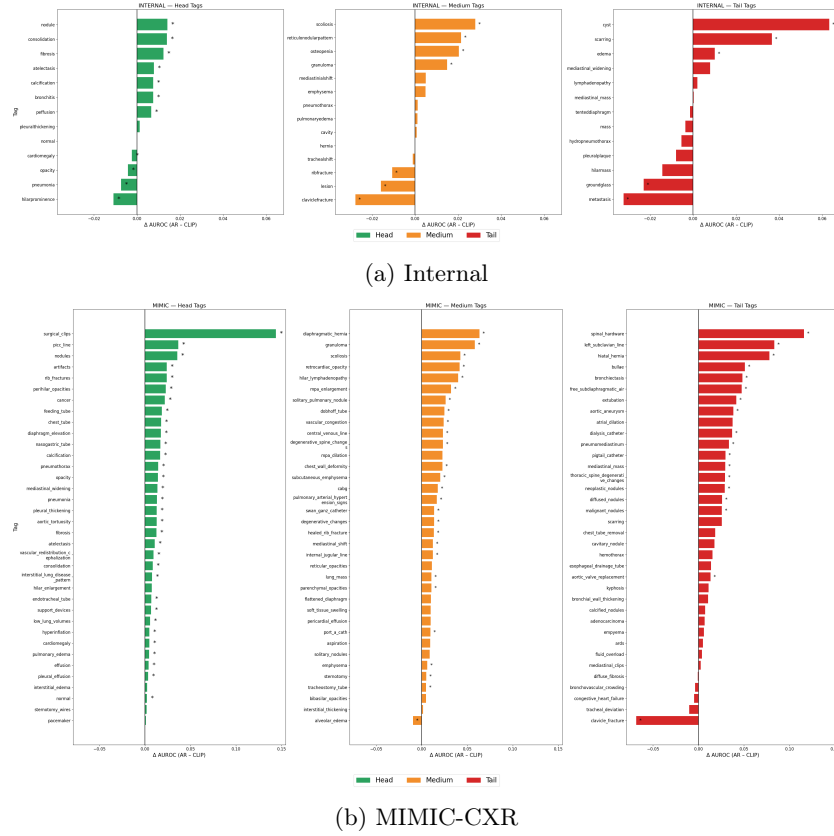


Fig. 7: Differences in AUROC, Δ , between autoregressive and Med-CLIP pre-training for (a) Internal and (b) MIMIC-CXR. Statistically significant differences under the DeLong test are marked with (*).

weaker than Med-AR-8B. A similar trend is observed on PadChest, where Med-AR-8B again provides the strongest uncertainty ranking, reducing EAURC from 0.314 to 0.217 overall, from 0.397 to 0.306 for head labels, from 0.351 to 0.240 for medium labels, and from 0.198 to 0.107 for tail labels. Med-AR-2B is consistently second-best on both datasets.

The CheXpert results add an important and somewhat unexpected nuance. Both autoregressive variants outperform Med-CLIP on EAURC across all prevalence groups, but here *Med-AR-2B*, not *Med-AR-8B*, achieves the best uncertainty ranking. Specifically, *Med-AR-2B* reduces EAURC from 0.212 (Med-CLIP) to 0.162 overall, from 0.406 to 0.356 for head labels, from 0.190 to 0.113 for medium labels, and from 0.040 to 0.018 for tail labels, with *Med-AR-8B* second-best throughout. This is particularly notable because *Med-AR-2B* has *lower* AUROC and AUPRC than Med-CLIP on CheXpert. In other words, stronger discrimination does not automatically translate into better uncertainty rank-



Fig. 8: Differences in AUROC, Δ , between autoregressive and Med-CLIP pre-training for (a) PadChest and (b) CheXpert. Statistically significant differences under the DeLong test are marked with (*).

ing: despite being a weaker classifier on CheXpert, Med-AR-2B produces confidence scores that rank its own errors more effectively than either Med-CLIP or Med-AR-8B. This dissociation between discrimination and selective-prediction behavior highlights the value of evaluating these two axes independently when comparing pretraining paradigms.

In contrast, the internal benchmark shows the opposite pattern. Here, Med-CLIP achieves the lowest EAURC overall and within each subgroup, with values of 0.014 overall, 0.036 for head labels, 0.004 for medium labels, and 0.002 for tail labels. Both Med-AR variants perform worse on this benchmark, with Med-AR-8B ahead of Med-AR-2B overall and for head labels, and Med-AR-2B ahead for medium and tail labels. Both remain above Med-CLIP in every group. This is consistent with the more mixed behavior already observed on the internal dataset in the previous subsections: although Med-AR remains competitive in discrim-

ination metrics, those gains do not translate into better confidence ordering in this more heterogeneous setting.

Taken together, the uncertainty-aware evaluation reinforces the broader narrative of the paper while adding two important qualifications. First, on all three public benchmarks, Med-AR-8B improves AUROC, AUPRC, and EAURC relative to Med-CLIP under the evaluated protocol; however, the optimal AR variant for uncertainty ranking is not always the one that maximizes discrimination, as clearly demonstrated on CheXpert. Second, on the internal dataset Med-CLIP retains a clear advantage in EAURC, suggesting that stronger representation transfer and better uncertainty ordering need not always coincide. These observations make uncertainty-aware evaluation an essential complementary axis for comparing pretraining paradigms in clinical imaging.

Table 5: Excess Area Under the Risk–Coverage Curve (EAURC) for Med-CLIP, Med-AR-8B, and Med-AR-2B across datasets. Values are averaged over all labels (Overall) and separately over head, medium, and tail labels. Lower values indicate better uncertainty ranking. The best value in each row is shown in bold, and the second-best value is underlined.

Dataset	Group	Med-CLIP EAURC	Med-AR-8B (ours) EAURC	Med-AR-2B (ours) EAURC
Internal	Overall	0.014	<u>0.070</u>	0.073
	Head	0.036	<u>0.173</u>	0.194
	Medium	0.004	0.018	<u>0.016</u>
	Tail	0.002	0.022	<u>0.013</u>
MIMIC	Overall	0.196	0.169	<u>0.179</u>
	Head	0.365	0.334	<u>0.345</u>
	Medium	0.186	0.146	<u>0.165</u>
	Tail	0.038	0.027	<u>0.028</u>
PadChest	Overall	0.314	0.217	<u>0.274</u>
	Head	0.397	0.306	<u>0.358</u>
	Medium	0.351	0.240	<u>0.315</u>
	Tail	0.198	0.107	<u>0.153</u>
CheXpert	Overall	0.212	<u>0.191</u>	0.162
	Head	0.406	<u>0.384</u>	0.356
	Medium	0.190	<u>0.159</u>	0.113
	Tail	0.040	<u>0.031</u>	0.018

5.5 Comparison of pretrained encoders

The uncertainty-aware evaluation above compares Med-AR directly against Med-CLIP. We next broaden that comparison to ask whether the observed gains persist when the image encoder is initialized from a wider set of representation learning strategies, while keeping the downstream classifier fixed. To this end, we replace the pretrained encoder with alternatives spanning self-supervised, supervised, biomedical contrastive, knowledge-enhanced, and radiology-specific vision–language pretraining, and fine-tune all of them under the same ML-Decoder downstream setup on CheXpert, MIMIC-CXR, and PadChest.

Specifically, we compare RAD-DINO, an image-only self-supervised medical encoder trained beyond text supervision [32]; ARK (Ark⁺), a supervised chest radiography foundation model trained from heterogeneous expert-labeled datasets without manual label consolidation [28,27,29]; BiomedCLIP, a large-scale biomedical contrastive model pretrained on scientific image–text pairs [53]; MedKLIP, a knowledge-enhanced radiology vision–language pretraining method that aligns medical entities with image patches [47]; the EVA-based encoder (EVA-Base in the tables); BioViL-T, a radiology-specific contrastive vision–language model from the BiomedVLP family that builds on the CXR-BERT text encoder and improved domain-specific textual semantics [3]; CheXFound, a self-supervised chest X-ray foundation model pretrained on a large curated CXR corpus and originally paired with a global–local representation integration strategy for downstream adaptation [50]; and our proposed Med-AR-2B and Med-AR-8B encoders. Table 6 summarizes the resulting performance across the three public benchmarks. Label-wise results are organized by dataset in Supplementary Sections S3–S6.

Table 6 shows that the autoregressive encoders remain the strongest representation learners under a fixed downstream setup, and this conclusion now holds across all three public datasets. On CheXpert, Med-AR-8B achieves the best mAUROC and mAUPRC for the head, medium, and tail subsets, with MedCLIP consistently second-best and ARK/CheXFound close behind on the head subset. On MIMIC-CXR, Med-AR-8B is again best across all prevalence groups, while Med-AR-2B is second-best throughout, clearly ahead of the non-autoregressive baselines. On PadChest, Med-AR-2B is the strongest model across head, medium, and tail subsets, with Med-AR-8B second-best and ARK/CheXFound leading among the non-autoregressive alternatives. Among the non-AR baselines, ARK and CheXFound are the most competitive on the head and tail groups, whereas the other radiology-specific and biomedical baselines, including MedKLIP, EVA-Base, and BioViL-T, remain clearly below the best Med-AR variants on both ranking-based metrics for every dataset.

Table 6: Comparison of vision encoders initialized from different methods across CheXpert, MIMIC-CXR, and PadChest using mean AUROC and mean AUPRC. All encoders are fine-tuned with the same ML-Decoder head. Within each dataset, metric, and prevalence group, the best result is shown in bold and the second-best is underlined. H=head, M=medium, and T=tail.

Method	Grp	CheXpert		MIMIC-CXR		PadChest	
		mAUROC	mAUPRC	mAUROC	mAUPRC	mAUROC	mAUPRC
EVA-Base	H	0.7981	0.2918	0.8250	0.3013	0.7961	0.3463
	M	0.8223	0.0996	0.8115	0.0804	0.7934	0.2480
	T	0.7843	0.0333	0.7952	0.0326	0.8185	0.2581
RAD-DINO	H	0.7632	0.2473	0.8440	0.3287	0.7777	0.3109
	M	0.7847	0.0614	0.8380	0.1032	0.7764	0.2224
	T	0.7460	0.0193	0.8289	0.0397	0.8056	0.2406
ARK	H	0.8430	0.3680	0.8837	0.4202	0.8419	0.4384
	M	0.8593	0.1451	0.8809	0.1817	0.8510	0.3633
	T	0.8325	0.0649	0.8816	0.1111	0.9047	0.4747
CheXFound	H	0.8429	0.3692	0.8809	0.4098	0.8487	0.4558
	M	0.8670	0.1665	0.8849	0.1987	0.8584	0.3770
	T	0.8384	0.0815	0.8736	0.1008	0.8950	0.4476
BiomedCLIP	H	0.7392	0.2360	0.8582	0.3719	0.6630	0.2034
	M	0.7604	0.0715	0.8536	0.1542	0.6705	0.1471
	T	0.7311	0.0269	0.8425	0.0791	0.7241	0.1970
MedKLIP	H	0.7847	0.2786	0.8112	0.2839	0.8132	0.3812
	M	0.7980	0.0777	0.7866	0.0621	0.8172	0.3045
	T	0.7411	0.0201	0.7628	0.0298	0.8344	0.3406
BioViL-T	H	0.6604	0.1676	0.7117	0.1881	0.6626	0.1858
	M	0.6217	0.0200	0.6424	0.0207	0.6632	0.1510
	T	0.6113	0.0072	0.5833	0.0061	0.7104	0.2087
MedCLIP	H	<u>0.8433</u>	<u>0.3710</u>	0.8756	0.4010	0.8425	0.4428
	M	<u>0.8790</u>	<u>0.1802</u>	0.8841	0.1902	0.8531	0.3678
	T	<u>0.8526</u>	<u>0.0922</u>	0.8733	0.1033	0.9002	0.4762
Med-AR-2B (ours)	H	0.8310	0.3614	<u>0.8846</u>	<u>0.4264</u>	0.8571	0.4706
	M	0.8586	0.1682	<u>0.8881</u>	<u>0.2175</u>	0.8674	0.3941
	T	0.8365	0.0829	<u>0.8851</u>	<u>0.1185</u>	0.9143	0.5072
Med-AR-8B (ours)	H	0.8565	0.3978	0.8918	0.4432	<u>0.8497</u>	<u>0.4583</u>
	M	0.8920	0.2054	0.9028	0.2435	<u>0.8595</u>	<u>0.3818</u>
	T	0.8788	0.1151	0.8967	0.1441	<u>0.9060</u>	<u>0.4940</u>

This comparison extends the analysis beyond a single contrastive baseline under a common downstream head. It is not a component ablation: the pretrained systems differ in architecture, pretraining data, and optimization. The results support transfer performance under the evaluated setup, without attributing differences solely to the pretraining objective or excluding interactions with the downstream head.

5.6 Discussion

Taken together, the results support five main observations.

First, on all three public benchmarks—PadChest, MIMIC-CXR, and CheXpert—autoregressive pretraining improves mAUROC and mAUPRC over Med-CLIP with Med-AR-8B across all prevalence groups, with nominal statistical support in most comparisons, with the strongest and most consistent gains delivered by Med-AR-8B. On MIMIC-CXR and CheXpert, Med-AR-8B is best overall and in every prevalence subgroup; on PadChest, the smaller Med-AR-2B is best while Med-AR-8B remains a strong second-best. The consistency of this pattern across aggregate metrics, paired Wilcoxon tests, and label-wise DeLong comparisons suggests a broad improvement in representation transfer rather than an artifact of a few abnormalities. The tail-label gains are consistent with the hypothesis that structured and region-aware autoregressive supervision benefits rare-finding recognition, but do not directly demonstrate the proposed visual-grounding mechanism.

Second, the internal benchmark is more nuanced. Med-AR-2B shows a modest mAUROC edge with partial statistical support, but mAUPRC results are mixed, tail-label separation is not uniform, and Med-CLIP retains a clear EAURC advantage in every subgroup. The two paradigms are therefore best described as broadly comparable on internal data, suggesting that the benefit of autoregressive pretraining is partly conditioned by dataset properties such as reporting consistency, label construction quality, and institutional heterogeneity.

Third, the EAURC analysis reveals that stronger discrimination does not guarantee better confidence-based risk ordering. Med-AR improves EAURC on all three public benchmarks, but CheXpert provides a striking example where discrimination and selective-prediction behavior decouple: Med-AR-2B has lower AUROC and AUPRC than Med-CLIP yet achieves the best EAURC across every CheXpert subgroup. On the internal dataset, the opposite dissociation occurs, with Med-CLIP achieving the best EAURC despite Med-AR matching or exceeding it on discrimination metrics. Together, these findings underscore the importance of evaluating selective-prediction behavior as a distinct axis rather than assuming it follows discrimination gains.

Fourth, the encoder comparison across CheXpert, MIMIC-CXR, and PadChest shows that Med-AR encoders outperform not only Med-CLIP but also a broad set of self-supervised, supervised, biomedical contrastive, knowledge-enhanced, and radiology-specific VLP baselines under the same ML-Decoder head, supporting the effectiveness of the complete pretraining recipe under this downstream setup.

Finally, the preferred autoregressive variant is dataset-dependent: Med-AR-2B leads on Internal and PadChest, while Med-AR-8B leads on MIMIC-CXR and CheXpert. This is most consequential on CheXpert, where Med-AR-2B fails to outperform Med-CLIP on discrimination while Med-AR-8B does so decisively. This suggests an interaction between decoder capacity, the pretraining recipe, and the downstream task. The present experiments do not isolate the objective as the primary driver of the gains.

Overall, these findings identify radiology-native autoregressive pretraining as a strong foundation for long-tailed CXR classification on public benchmarks, while the mixed internal results highlight that representation superiority is not absolute and depends on both dataset characteristics and the evaluation axis under consideration.

5.7 Limitations

The comparison controls the downstream architecture, but the pretraining recipes differ in supervision and may differ in initialization and compute. No matched-supervision component ablation is available, so the results cannot establish a causal advantage of next-token prediction alone. The additional pretrained-encoder comparison also does not equalize architecture or pretraining data.

The expanded MIMIC-CXR and CheXpert labels are derived from reports. Inter-model concordance does not establish expert-level correctness, and unmentioned findings may be encoded as negatives. Generated reasoning and prominence targets may introduce additional noise. The 1,000-report concordance sample provides limited evidence for very rare labels, and the retained PadChest vocabulary excludes many less frequent findings. Each dataset is fine-tuned separately; these experiments do not constitute deployment on an unseen institution without adaptation.

The statistical analyses do not quantify training-seed variability or establish equivalence for non-significant comparisons. Label dependencies and multiple testing warrant caution in interpreting nominal significance. EAURC assesses selective prediction under the evaluated protocol rather than calibration or clinical utility. Independent expert validation, fuller reporting of data provenance and split construction, and prospective evaluation remain necessary to assess clinical applicability.

6 Conclusion and Future Work

We investigated whether the choice of vision-language pretraining recipe—contrastive versus autoregressive—affects downstream transfer for long-tailed multi-label chest X-ray classification. To reduce downstream architectural confounding, we compared MedCLIP-style contrastive pretraining and InternVL3-based autoregressive pretraining (Med-AR-2B and Med-AR-8B) using a shared InternViT vision architecture and a common ML-Decoder classification head,

and evaluated all models on an internal benchmark together with three public datasets: PadChest, MIMIC-CXR, and CheXpert.

Across the three public benchmarks, Med-AR-8B improves downstream AUROC and AUPRC over Med-CLIP in each prevalence group, with nominally significant paired per-label gains in most comparisons. Med-AR-8B performs best on MIMIC-CXR and CheXpert, while Med-AR-2B performs best on PadChest. The encoder comparison also favors Med-AR under the evaluated downstream setup. These observations support the complete autoregressive pretraining recipe, while the separate contributions of objective, supervision, and decoder scale remain unresolved.

On the internal benchmark, the two paradigms are better characterized as broadly comparable: Med-AR-2B shows a modest AUROC edge with partial statistical support, but AUPRC results are mixed and Med-CLIP retains a clear advantage in selective-prediction behavior as measured by EAURC. More broadly, our uncertainty-aware analysis uncovers a consequential decoupling between discrimination and confidence ranking. On CheXpert, Med-AR-2B achieves the best EAURC across all prevalence groups despite lower AUROC and AUPRC than Med-CLIP, while on the internal dataset the opposite dissociation occurs. Stronger discrimination therefore does not guarantee better confidence-based risk ordering, and prevalence-aware, uncertainty-aware evaluation should be treated as a first-class axis—rather than a secondary metric—when comparing medical vision–language pretraining strategies.

Several directions remain open. First, disentangling the contribution of the autoregressive objective itself from its richer supervision—structured multi-section report templates, abnormality-focused auxiliary outputs, and region-level annotations—will require component-wise ablations that we leave to future work. Second, dedicated strategies for rare-label learning, such as curriculum scheduling, meta-learning, and focused autoregressive pretraining on curated rare-finding corpora, may further close the gap on tail abnormalities and reduce residual variability across decoder scales. Third, integrating calibration techniques such as temperature scaling [26,48], conformal prediction [23,1], and post-hoc meta-models [39] would complement EAURC and support further assessment of reliability before clinical deployment. Finally, extending the LLM-based label-expansion framework to additional datasets and imaging modalities, and embedding the resulting confidence estimates into clinical workflows such as triage and radiologist–AI collaboration, are important next steps toward assessing real-world impact.

In summary, radiology-native autoregressive pretraining provides a strong foundation for long-tailed chest X-ray classification on public benchmarks, particularly for underrepresented findings. Its advantages, however, are not uniform across datasets or evaluation axes, and the Med-AR configuration that is best for discrimination need not coincide with the one that is best for selective prediction. These observations argue for a joint assessment of discrimination and uncertainty behavior when comparing medical vision–language pretraining paradigms, and

position radiology-native Med-AR as a starting point for further research on reliable chest X-ray classification.

7 Statements and Declarations

Author Contributions All authors contributed to the conception and design of the study. Janhavi Prabhu led the model training, methodology development, experimental design, and manuscript writing. Sahil contributed to model training, dataset preparation, label extraction and manuscript writing. Akshay Valsaraj contributed to implementation support and model development. Manoj Tadepalli and Preetham Putha provided oversight, technical guidance, and critical revisions of the manuscript. All authors reviewed and approved the final version of the manuscript.

Data Availability We plan to release the derived MIMIC-CXR and CheXpert label annotations and split identifiers following publication, subject to the source datasets' access and licensing requirements. Original images and reports remain subject to their respective access conditions.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Angelopoulos, A., Bates, S., Malik, J., Jordan, M.I.: Uncertainty sets for image classifiers using conformal prediction (Sep 2020)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-VL technical report (Feb 2025)
3. Bannur, S., Hyland, S., Liu, Q., Perez-Garcia, F., Ilse, M., Castro, D.C., Boecking, B., Sharma, H., Bouzid, K., Thieme, A., Schwaighofer, A., Wetscherek, M., Lungren, M.P., Nori, A., Alvarez-Valle, J., Oktay, O.: Learning to Exploit Temporal Structure for Biomedical Vision-Language Processing . In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15016–15027. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2023). <https://doi.org/10.1109/CVPR52729.2023.01442>, <https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.01442>
4. Boecking, B., Usuyama, N., Bannur, S., Castro, D.C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., et al.: Making the most of text semantics to improve biomedical vision–language processing. In: European conference on computer vision. pp. 1–21. Springer (2022)
5. Bustos, A., Pertusa, A., Salinas, J.M., de la Iglesia-Vayá, M.: Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis* **66**, 101797 (2019), <https://api.semanticscholar.org/CorpusID:58981612>
6. Chen, J., Yang, J., Wu, H., Li, D., Gao, J., Zhou, T., Xiao, B.: Florence-VL: Enhancing vision-language models with generative vision encoder and depth-breadth fusion (Dec 2024)

7. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24185–24198 (2024)
8. DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L.: Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* **44**(3), 837–845 (1988). <https://doi.org/10.2307/2531595>, <https://pubmed.ncbi.nlm.nih.gov/3203132/>
9. Desai, K., Johnson, J.: Virtex: Learning visual representations from textual annotations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11162–11173 (June 2021)
10. El-Nouby, A., Klein, M., Zhai, S., Bautista, M.Á., Shankar, V., Toshev, A.T., Susskind, J.M., Joulin, A.: Scalable pre-training of large autoregressive image models. In: Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 235, pp. 12371–12384. PMLR (2024)
11. Faghani, S., Moassefi, M., Rouzrokh, P., Khosravi, B., Baffour, F.I., Ringler, M.D., Erickson, B.J.: Quantifying uncertainty in deep learning of radiologic images. *Radiology* **308**(2), e222217 (2023)
12. Fini, E., Shukor, M., Li, X., Dufter, P., Klein, M., Haldimann, D., Aitharaju, S., da Costa, V.G.T., Béthune, L., Gan, Z., Toshev, A., Eichner, M., Nabi, M., Yang, Y., Susskind, J., El-Nouby, A.: Multimodal autoregressive pre-training of large vision encoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9641–9654 (June 2025)
13. Geifman, Y., El-Yaniv, R.: Selectivenet: A deep neural network with an integrated reject option. In: International Conference on Machine Learning (2019), <https://api.semanticscholar.org/CorpusID:59316904>
14. Geifman, Y., Uziel, G., El-Yaniv, R.: Bias-reduced uncertainty estimation for deep neural classifiers. In: International Conference on Learning Representations (2018), <https://api.semanticscholar.org/CorpusID:52901777>
15. Holste, G., Wang, S., Jiang, Z., Shen, T.C., Shih, G., Summers, R.M., Peng, Y., Wang, Z.: Long-tailed classification of thorax diseases on chest x-ray: A new benchmark study. In: MICCAI Workshop on Data Augmentation, Labelling, and Imperfections. pp. 22–32. Springer (2022)
16. Holste, G., Zhou, Y., Wang, S., Jaiswal, A., Lin, M., Zhuge, S., Yang, Y., Kim, D., Nguyen-Mau, T.H., Tran, M.T., et al.: Towards long-tailed, multi-label disease classification from chest x-ray: Overview of the cxr-lt challenge. *Medical Image Analysis* p. 103224 (2024)
17. Huang, S.C., Shen, L., Lungren, M.P., Yeung, S.: GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE (Oct 2021)
18. Hüllermeier, E., Waegeman, W.: Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach. Learn.* **110**(3), 457–506 (Mar 2021)
19. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghighi, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D.A., Halabi, S.S., Sandberg, J.K., Jones, R., Larson, D.B., Langlotz, C.P., Patel, B.N., Lungren, M.P., Ng, A.Y.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison (2019), <https://arxiv.org/abs/1901.07031>

20. Jing, D., He, X., Luo, Y., Fei, N., Yang, G., Wei, W., Zhao, H., Lu, Z.: Fineclip: Self-distilled region-based clip for better fine-grained understanding. In: *Advances in Neural Information Processing Systems* (2024)
21. Johnson, A., Pollard, T., Mark, R., Berkowitz, S., Horng, S.: MIMIC-CXR Database. PhysioNet (Jul 2024). <https://doi.org/10.13026/4jqj-jw95>, <https://doi.org/10.13026/4jqj-jw95>, version 2.1.0
22. Johnson, A.E.W., Pollard, T., Mark, R., Berkowitz, S., Horng, S.: The mimic-cxr database (2019). <https://doi.org/10.13026/C2JT1Q>, <https://physionet.org/content/mimic-cxr/>
23. Karimi, H., Samavi, R.: Quantifying deep learning model uncertainty in conformal prediction. Toronto Metropolitan University (Jan 2024)
24. Kim, D.: Chexfusion: Effective fusion of multi-view features using transformers for long-tailed chest x-ray classification. 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW) pp. 2694–2702 (2023), <https://api.semanticscholar.org/CorpusID:260704332>
25. Kimi Team, Du, A., Yin, B., Xing, B., Qu, B., Wang, B., Chen, C., Zhang, C., Du, C., Wei, C., Wang, C., Zhang, D., Du, D., Wang, D., Yuan, E., Lu, E., Li, F., Sung, F., Wei, G., Lai, G., Zhu, H., Ding, H., Hu, H., Yang, H., Zhang, H., Wu, H., Yao, H., Lu, H., Wang, H., Gao, H., Zheng, H., Li, J., Su, J., Wang, J., Deng, J., Qiu, J., Xie, J., Wang, J., Liu, J., Yan, J., Ouyang, K., Chen, L., Sui, L., Yu, L., Dong, M., Dong, M., Xu, N., Cheng, P., Gu, Q., Zhou, R., Liu, S., Cao, S., Yu, T., Song, T., Bai, T., Song, W., He, W., Huang, W., Xu, W., Yuan, X., Yao, X., Wu, X., Li, X., Zu, X., Zhou, X., Wang, X., Charles, Y., Zhong, Y., Li, Y., Hu, Y., Chen, Y., Wang, Y., Liu, Y., Miao, Y., Qin, Y., Chen, Y., Bao, Y., Wang, Y., Kang, Y., Liu, Y., Dong, Y., Du, Y., Wu, Y., Wang, Y., Yan, Y., Zhou, Z., Li, Z., Jiang, Z., Zhang, Z., Yang, Z., Huang, Z., Huang, Z., Zhao, Z., Chen, Z., Lin, Z.: Kimi-VL technical report (Jun 2025)
26. Kull, M., Perello-Nieto, M., Kängsepp, M., Filho, T.S., Song, H., Flach, P.: Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with dirichlet calibration (Oct 2019)
27. Ma, D., Pang, J., Gotway, M.B., Liang, J.: Foundation ark: Accruing and reusing knowledge for superior and robust performance. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. pp. 651–662. Springer Nature Switzerland, Cham (2023)
28. Ma, D., Pang, J., Gotway, M.B., Liang, J.: A fully open AI foundation model applied to chest radiography. *Nature* pp. 1–11 (2025)
29. Ma, D., Pang, J., Senthil Velan, S., Gotway, M.B., Liang, J.: Ark+: Supervised training a single high-performance ai foundation model from many differently labeled datasets—no label consolidation required. *Medical Image Analysis* **108**, 103828 (2026). <https://doi.org/10.1016/j.media.2025.103828>, <https://www.sciencedirect.com/science/article/pii/S1361841525003743>
30. Moon, J.H., Lee, H., Shin, W., Kim, Y.H., Choi, E.: Multi-modal understanding and generation for medical images and text via vision-language pre-training. *IEEE Journal of Biomedical and Health Informatics* **26**(12), 6070–6080 (2022)
31. Park, W., Park, I., Kim, S., Ryu, J.B.: Robust asymmetric loss for multi-label long-tailed learning. 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW) pp. 2703–2712 (2023), <https://api.semanticscholar.org/CorpusID:260775967>
32. Pérez-García, F., Sharma, H., Bond-Taylor, S., Bouzid, K., Salvatelli, V., Ilse, M., Bannur, S., Castro, D.C., Schwaighofer, A., Lungren, M.P., Wetscherek, M.T.,

- Codella, N., Hyland, S.L., Alvarez-Valle, J., Oktay, O.: Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence* **7**, 119–130 (2025). <https://doi.org/10.1038/s42256-024-00965-w>
33. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *International conference on machine learning*. pp. 8748–8763. PmLR (2021)
 34. Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., Lu, J.: Denseclip: Language-guided dense prediction with context-aware prompting. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 18082–18091 (June 2022)
 35. Ridnik, T., Sharir, G., Ben-Cohen, A., Ben-Baruch, E., Noy, A.: Ml-decoder: Scalable and versatile classification head. *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* pp. 32–41 (2021), <https://api.semanticscholar.org/CorpusID:244709117>
 36. Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M., Zelnik-Manor, L.: Asymmetric loss for multi-label classification. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 82–91 (2021)
 37. Saito, T., Rehmsmeier, M.: The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE* **10**(3), e0118432 (2015). <https://doi.org/10.1371/journal.pone.0118432>, <https://pubmed.ncbi.nlm.nih.gov/25738806/>
 38. Sensoy, M., Saleki, M., Julier, S., Aydogan, R., Reid, J.: Misclassification risk and uncertainty quantification in deep classifiers. In: *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. vol. 86, pp. 2483–2491. IEEE (Jan 2021)
 39. Shen, M., Bu, Y., Sattigeri, P., Ghosh, S., Das, S., Wornell, G.: Post-hoc uncertainty learning using a dirichlet meta-model. *Proc. Conf. AAAI Artif. Intell.* **37**(8), 9772–9781 (Jun 2023)
 40. Sun, Q., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, Y., Gao, H., Liu, J., Huang, T., Wang, X.: Emu: Generative pretraining in multimodality. In: *The Twelfth International Conference on Learning Representations* (2024)
 41. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing* **2022**, 3876–3887 (2022), <https://api.semanticscholar.org/CorpusID:252992913>
 42. Wang, Z., Yu, J., Yu, A.W., Dai, Z., Tsvetkov, Y., Cao, Y.: Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904* (2021)
 43. Wei, Y., Wang, X., Ong, H., Zhou, Y., Flanders, A., Shih, G., Peng, Y.: Enhancing disease detection in radiology reports through fine-tuning lightweight llm on weak labels. *arXiv preprint arXiv:2409.16563* (2024)
 44. Whata, A., Dibeco, K., Madzima, K., Obagbuwa, I.: Uncertainty quantification in multi-class image classification using chest x-ray images of covid-19 and pneumonia. *Frontiers in Artificial Intelligence* **7**, 1410841 (2024)
 45. Wiehe, A., Schneider, F., Blank, S., Wang, X., Zorn, H.P., Biemann, C.: Language over labels: Contrastive language supervision exceeds purely label-supervised classification performance on chest x-rays. In: *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th*

- International Joint Conference on Natural Language Processing: Student Research Workshop. pp. 76–83. Association for Computational Linguistics, Stroudsburg, PA, USA (2022)
46. Wilcoxon, F.: Individual comparisons by ranking methods. *Biometrics Bulletin* **1**(6), 80–83 (1945). <https://doi.org/10.2307/3001968>
 47. Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Medkclip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 21315–21326 (2023). <https://doi.org/10.1109/ICCV51070.2023.01954>
 48. Xie, J., Chen, A.S., Lee, Y., Mitchell, E., Finn, C.: Calibrating language models with adaptive temperature scaling (Sep 2024)
 49. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025), <https://arxiv.org/abs/2505.09388>
 50. Yang, Z., Xu, X., Zhang, J., Wang, G., Kalra, M.K., Yan, P.: Chest x-ray foundation model with global and local representations integration. *IEEE Transactions on Medical Imaging* **44**(12), 4787–4799 (2025). <https://doi.org/10.1109/TMI.2025.3581907>
 51. You, K., Gu, J., Ham, J., Park, B., Kim, J., Hong, E.K., Baek, W., Roh, B.: Cxr-clip: Toward large scale chest x-ray language-image pre-training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 101–111. Springer (2023)
 52. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research* (2022)
 53. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., Tupini, A., Wang, Y., Mazzola, M., Shukla, S., Liden, L., Gao, J., Lungren, M.P., Naumann, T., Wang, S., Poon, H.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023). <https://doi.org/10.48550/arXiv.2303.00915>
 54. Zhao, Y., Chen, S., Chen, Q., Hu, Z.: Combining loss reweighting and sample resampling for long-tailed instance segmentation. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (2023). <https://doi.org/10.1109/ICASSP49357.2023.10094303>
 55. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., Gao, Z., Cui, E., Wang, X., Cao, Y., Liu, Y., Wei, X., Zhang, H., Wang, H., Xu, W., Li, H., Wang, J., Deng, N., Li, S., He, Y., Jiang, T., Luo, J., Wang, Y., He, C., Shi, B., Zhang, X., Shao, W., He, J., Xiong, Y., Qu, W., Sun, P., Jiao, P., Lv, H., Wu, L., Zhang, K., Deng, H., Ge, J., Chen, K., Wang, L., Dou, M., Lu, L., Zhu, X., Lu, T., Lin, D., Qiao, Y., Dai, J., Wang, W.: InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models (Apr 2025)
 56. Zou, K., Chen, Z., Yuan, X., Shen, X., Wang, M., Fu, H.: A review of uncertainty estimation and its application in medical imaging (Feb 2023)

Supplementary Material

Med-AR: Autoregressive Vision-Language Pretraining for Long-Tailed Chest X-Ray Classification and Uncertainty-Aware Evaluation

This supplement provides label-construction details and the existing per-label results. Results are grouped by dataset so that discrimination and operating-point metrics can be inspected together. In the tables, * denotes tail labels and † denotes medium-frequency labels; unmarked labels are head labels. Bold and underlined entries indicate the highest and second-highest distinct displayed values, respectively, with ties sharing rank. These marks are descriptive, not significance tests.

Section Contents	
S1	Target preparation and label concordance
S2	Cross-dataset summary and operating points
S3	Internal: consolidated per-label metrics
S4	CheXpert: all five per-label metrics
S5	MIMIC-CXR: all five per-label metrics
S6	PadChest: all five per-label metrics

S1 Target Preparation and Label Concordance

S1.1 Pretraining Target Structure

The autoregressive targets use 15 anatomical subheadings with finding-specific fields, together with Findings, Impression, Summary, Reasoning, and Prominence Score sections. Prominence scores range from 0 to 5 and are generated from report text. Qwen3 is used to construct these report-derived targets; expert bounding-box annotations provide separate region-level supervision for approximately 20 findings. Non-mention is mapped to negation in the reported template construction, which can introduce false-negative supervision. The reasoning and prominence fields are synthetic targets, not independent expert image annotations. This description summarizes the target structure and does not specify an exact reproducible prompt.

S1.2 LLM-Based Label Expansion Pipeline

Parsing and construction of the presence matrix. Each LLM response was parsed from its hierarchical JSON representation into tuples of (**parent group**, **child tag**, **presence**). Although the extraction schema additionally permitted attributes such as size, location, and characteristics, only the binary **presence** field was used for label construction and concordance analysis. Child labels were indexed jointly with their parent group to avoid collisions between identically named labels occurring in different sections of the schema.

For a successfully parsed report, a tag explicitly marked as present was assigned a value of 1. Tags belonging to the model vocabulary but not emitted for that report were treated as implicitly absent (0). In contrast, reports for which the JSON response could not be parsed were treated as missing rather than negative and were excluded from the corresponding comparisons. This distinction prevented parsing failures from being interpreted as evidence for absence. A zero denotes the extraction convention and does not independently establish image-level absence.

Vocabulary normalization. The predefined schema was supplemented with an open-ended **other_findings** field to preserve findings not anticipated by the original taxonomy. Consequently, the raw outputs did not share a perfectly fixed vocabulary. In the full concordance analysis, the seven models collectively produced 5,586 distinct keys, whereas only 193 canonical tags were emitted by all seven models; most of the remaining variation originated from **other_findings**. This expansion primarily reflected lexical fragmentation—for example, singular/plural forms such as **chest_tube** and **chest_tubes**—rather than thousands of distinct radiographic concepts.

We therefore normalized the extracted vocabulary before constructing the final label space. Formatting variants and alternative representations of the same concept were first standardized, after which semantically equivalent labels were mapped to a canonical label. Importantly, normalization was performed at the level of clinical concepts rather than by string similarity alone. Labels were therefore merged only when they represented the same finding or a difference in naming granularity, while clinically related but distinct findings remained separate. Mapping to a parent group for concordance analysis should not be interpreted as evidence that the child concepts are interchangeable.

Synonym and hierarchical mapping. In the absence of ground-truth annotations for the expanded label set, concordance across multiple LLMs was used as a check of inter-model reproducibility. Agreement among models was interpreted as supporting evidence that the extracted labels were reproducible rather than as a mechanism for defining the label mappings themselves. The concordance analysis also provided insight into the nature of residual disagreement. In particular, many apparent discrepancies occurred between synonymous or hierarchically related labels. For example, when `solitary_pulmonary_nodule` was not emitted, `nodules` was emitted in 96% of such cases; similarly, `pleural_effusion` was emitted in 95% of cases in which `effusion` was absent. These observations indicate that a substantial fraction of tag-level disagreement reflected differences in terminology or label granularity rather than failure to identify the underlying radiographic finding, and informed inspection of the normalization and mapping procedure without establishing clinical correctness.

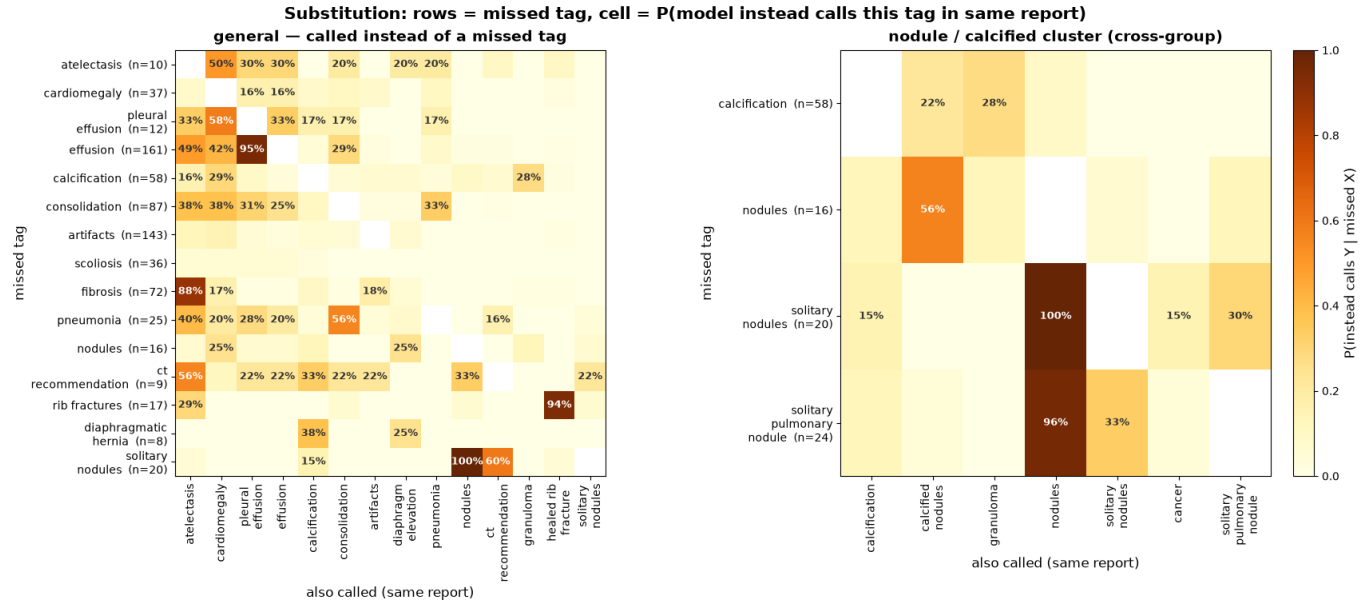


Fig. S1: **Substitution patterns among discordant labels.** For each missed tag (rows), cells show the probability that the same model assigned an alternative tag (columns) in the same report. High substitution rates identify cases where apparent disagreement reflects alternative label routing rather than failure to detect the underlying finding, as observed for `effusion/pleural_effusion` and the nodule family.

We consequently examined related label families for normalization and parent-level concordance analysis. These families include both synonyms and related but non-equivalent findings; parent-level agreement does not establish equivalence at the child-label level. These included the nodule family; `effusion/pleural_effusion`; low-lung-volume variants; COPD-related terminology such as hyperinflation and emphysema; ILD/fibrotic terminology; and closely related edema terminology. The mapping was deliberately conservative. In particular, `nodule`, `mass`, and `opacity` were maintained as separate concepts despite their potential co-occurrence, since substitutions among these categories can represent genuine differences in radiographic interpretation rather than lexical variation. Similarly, the airspace-opacity family (e.g., consolidation, pneumonia, perihilar opacity, and edema) was not indiscriminately collapsed, as the concordance analysis identified this group as one of the remaining sources of substantive clinical disagreement.

The effect of synonym routing was also assessed by comparing agreement at different levels of the hierarchy. Collapsing child labels into their parent finding groups increased mean pairwise Cohen's κ from 0.83 to 0.87, indicating that a substantial component of child-level disagreement disappeared when differences in terminology and granularity were removed. Thus, parent-level agreement was used as an additional diagnostic for distinguishing genuine disagreement from alternative routing within the taxonomy.

Handling free-form and ambiguous labels. Free-form `other_findings` labels received additional scrutiny because their vocabulary was substantially less standardized. Device and incidental findings were particularly affected: labels such as central venous catheters, PICC lines, chest tubes, surgical clips, and degenerative changes exhibited low apparent tag-level agreement despite frequently representing differences in naming or schema coverage. Free-form keys were

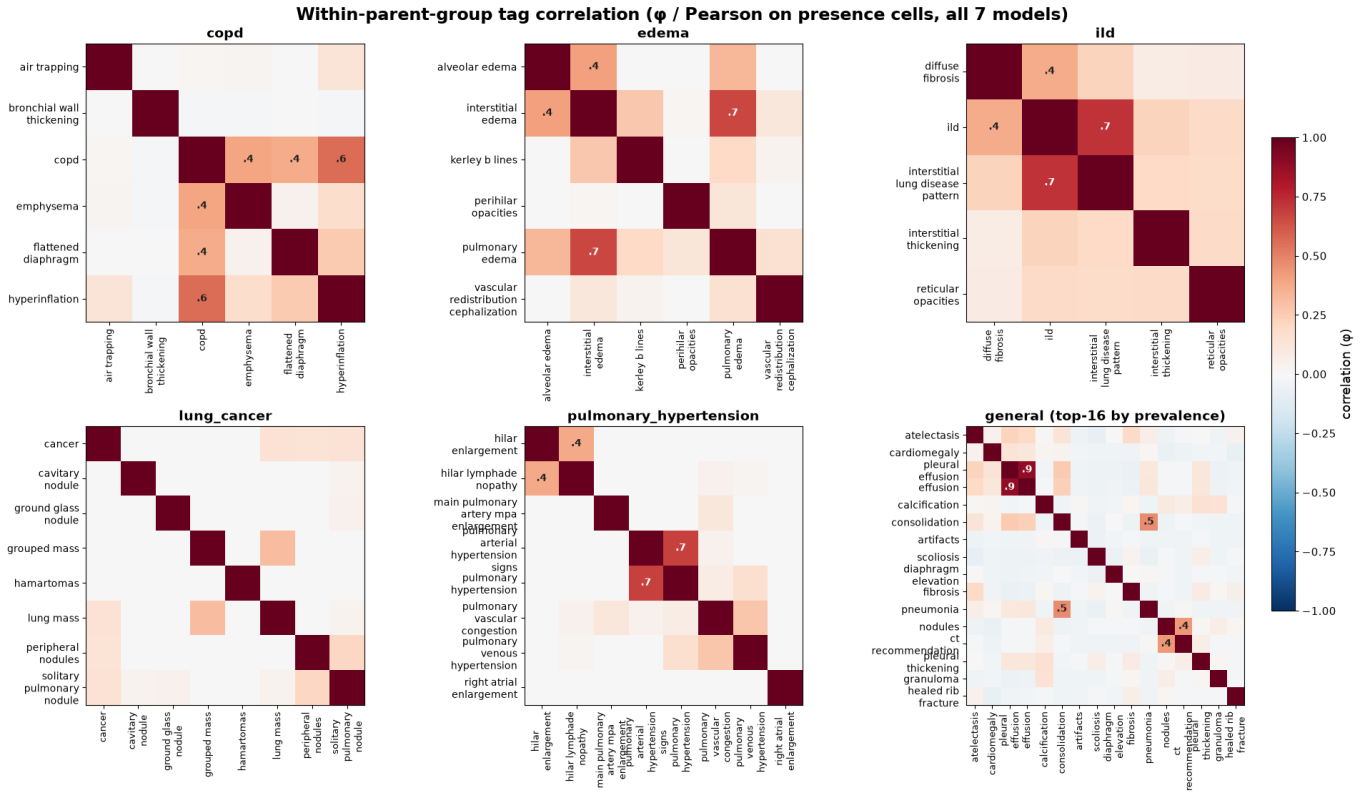


Fig. S2: **Within-group correlation of extracted labels across the seven LLMs.** Pairwise ϕ correlations between binary tag-presence indicators were computed across pooled model-report observations within each parent group. Higher ϕ indicates greater co-occurrence, capturing related, overlapping, or potentially synonymous labels.

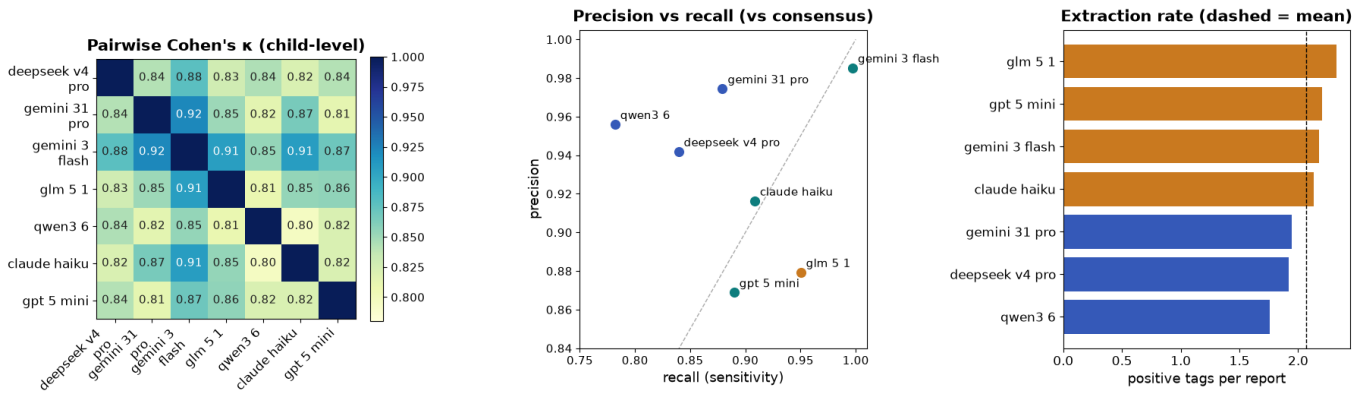


Fig. S3: **Inter-model agreement and extraction characteristics across seven LLMs.** **Left:** Pairwise Cohen's κ at the child-label level, demonstrating consistently high agreement between models. **Middle:** Precision–recall characteristics relative to the seven-model majority consensus, illustrating differences in model extraction behavior from conservative (higher precision, lower recall) to more inclusive extraction. **Right:** Mean number of positive tags extracted per report for each model; the dashed line indicates the seven-model mean.

therefore not automatically promoted to independent classes. Instead, recurrent keys were mapped to an existing canonical concept where possible; lexical variants were consolidated; and poorly supported or non-reproducible keys were flagged for taxonomy review.

The **artifacts** label was handled separately because its disagreement pattern differed from synonym routing. Only 92 reports reached majority consensus for **artifacts**, although at least one model emitted the label in 228 reports. Moreover, models emitting **artifacts** generally also emitted the corresponding radiographic abnormalities, indicating that **artifacts** behaved as an additional, model-dependent catch-all rather than as an alternative name for another finding. Such labels were therefore not merged with co-occurring clinical findings and were instead treated as candidates for stricter definition or removal.

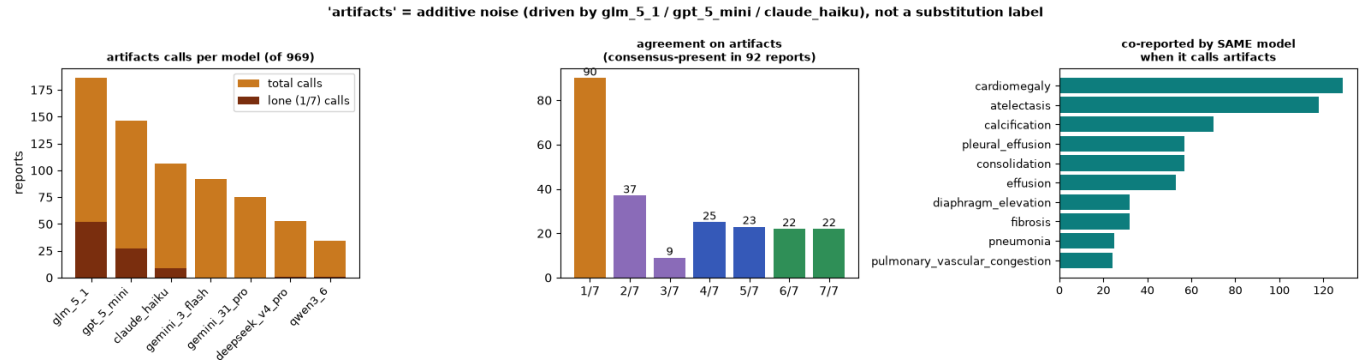


Fig. S4: **Model-specific behavior of the artifacts label.** The **artifacts** label shows substantial model-specific variation and is particularly frequent for GLM 5.1. This reflects a model-specific labeling tendency in which GLM 5.1 often assigns the generic **artifacts** label alongside more specific findings, such as devices or tubes. The high frequency of single-model calls consequently suggests that **artifacts** represents an inconsistently interpreted label rather than systematic disagreement in the underlying radiographic findings.

Taxonomy auditing and filtering. Following normalization, tags were audited using several complementary indicators of instability. These included the number of models containing the tag in their vocabulary, prevalence across models, pairwise agreement and Cohen's κ , frequency of single-model positive calls, and the extent to which disagreement disappeared at the parent-group level. Tags occurring in only a minority of model vocabularies were interpreted primarily as schema-coverage differences rather than reliable independent categories.

Finally, majority voting across the seven models was used during concordance analysis to identify systematic routing and extraction differences. A report-tag pair was considered consensus-positive when at least four of seven models marked it as present. This consensus served as an analytical reference rather than ground truth: over-calls and under-calls were defined relative to the $\geq 4/7$ majority and were subsequently examined for systematic substitutions. This procedure was particularly useful for separating consensus-relative omissions from vocabulary routing; for example, the apparently frequent under-calling of **effusion** was largely attributable to models using **pleural_effusion** instead. Overall, the post-processing procedure was designed to reduce vocabulary fragmentation without artificially increasing agreement by collapsing clinically distinct findings.

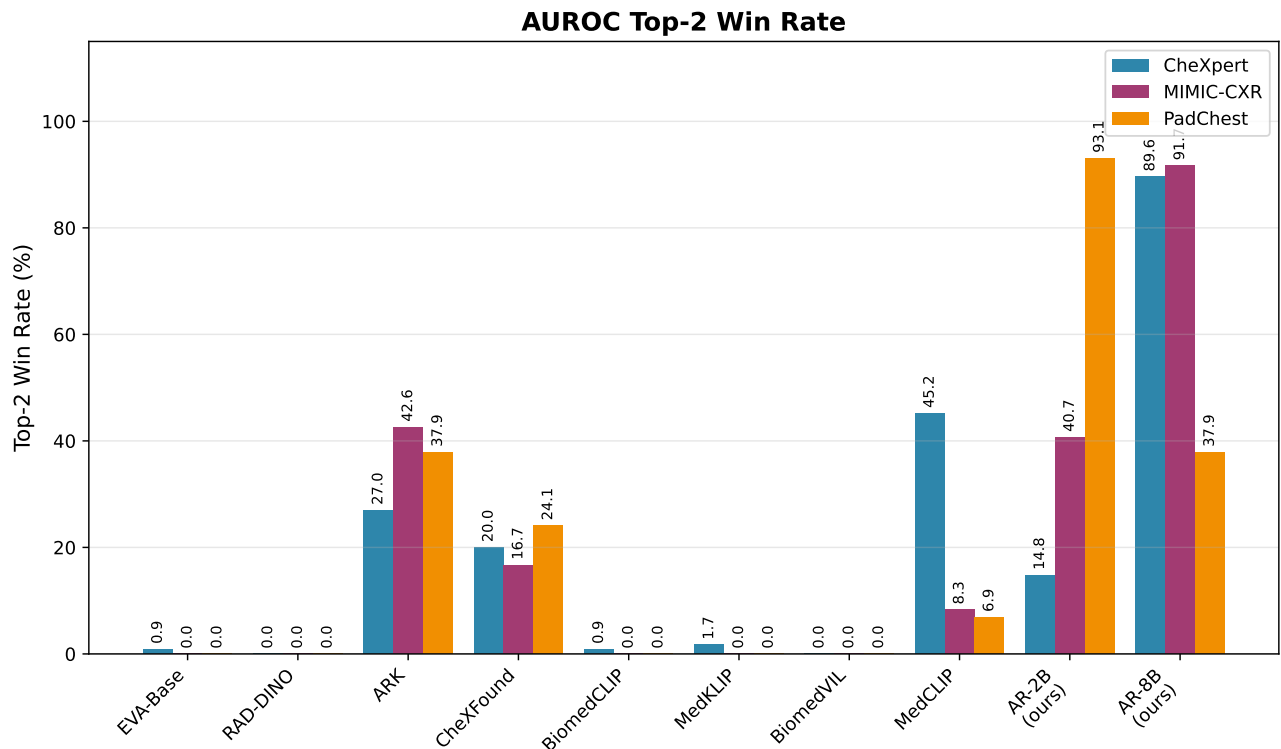


Fig.S5: Top-2 Win Rate (%) under AUROC across CheXpert, MIMIC-CXR, and PadChest. Our Med-AR-2B and Med-AR-8B models achieve top-2 placements on a larger fraction of tags than prior encoders. Numeric values are annotated above each bar for clarity.

S2 Cross-Dataset Summary and Operating Points

S2.1 Operating-Point Metrics

For public-dataset sensitivity, specificity, and Youden’s J , thresholds are selected per label by maximizing $J = \text{Sensitivity} + \text{Specificity} - 1$ on the validation set. These operating points summarize the sensitivity–specificity trade-off, rather than a clinically validated deployment threshold. The internal per-label table is described in the source analysis as validation-set performance with class-wise J -maximizing thresholds; it should be distinguished from the held-out aggregate comparison in the main text.

S2.2 Top-2 Win Rate Analysis for Public Datasets

To provide a holistic view of model competitiveness across the entire label space of chest radiograph benchmarks, we summarize the per-tag performance of every method using a *Top-2 Win Rate* metric. For each tag in a given dataset, we rank all ten competing encoders by their score and award a “win” to any model that appears among the top two. The win rate of a model is then defined as the fraction of tags (expressed as a percentage) on which it achieves a top-2 placement. This formulation rewards models that are *consistently* strong across the full tag distribution, rather than those that excel on only a handful of well-represented findings.

Figure S5 reports the Top-2 Win Rate measured under AUROC, while Figure S6 reports the same analysis under AUPRC. We evaluate on three widely used chest X-ray benchmarks CheXpert, MIMIC-CXR, and PadChest covering a broad spectrum of pathology labels ranging from common findings to long-tailed diagnoses. Med-AR-8B has the highest reported top-2 rates on CheXpert (89.6% AUROC, 85.2% AUPRC) and MIMIC-CXR (91.7% AUROC, 78.7% AUPRC), while Med-AR-2B leads on PadChest (93.1% AUROC, 96.6% AUPRC). This summary describes relative placement across labels; it does not measure effect size or establish rare-label clinical performance.

Per-label values are provided by dataset in Sections S4–S6.

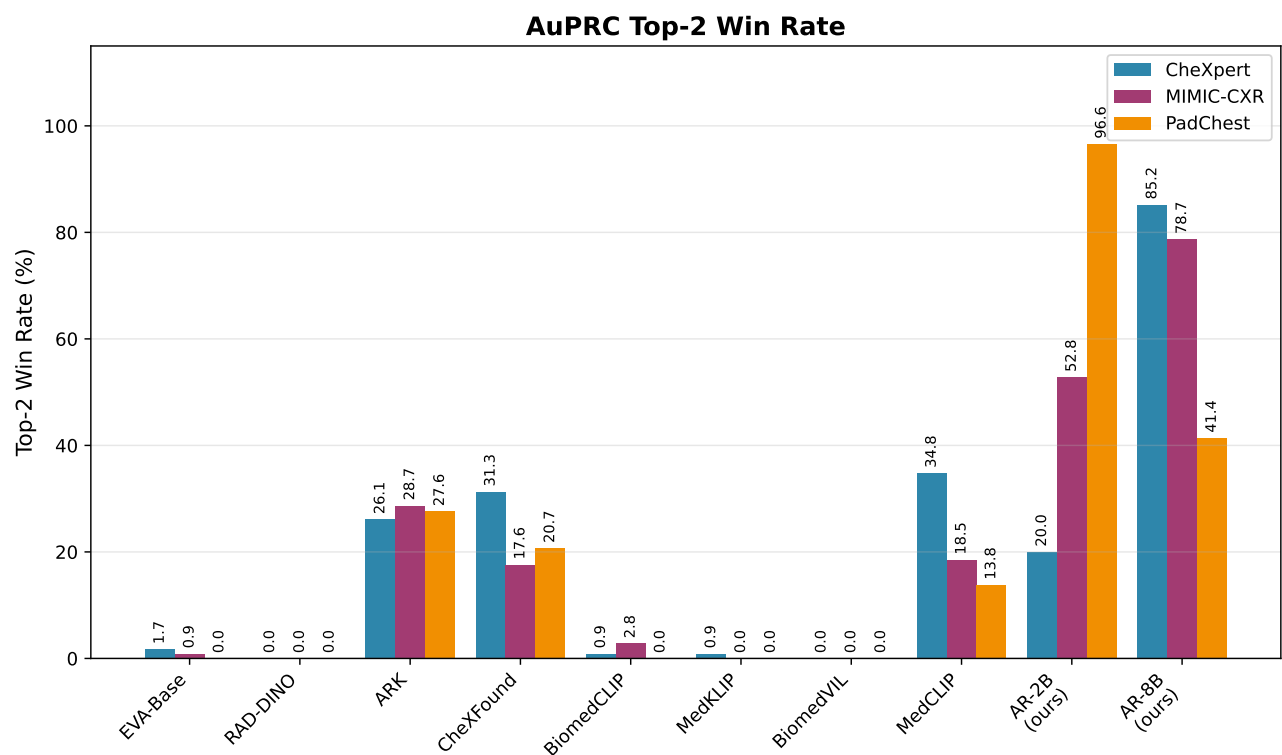


Fig. S6: Top-2 Win Rate (%) under AUPRC across CheXpert, MIMIC-CXR, and PadChest. Consistent with the AUROC results, our proposed models lead on precision-recall performance, with Med-AR-2B particularly excelling on the long-tailed PadChest benchmark.

S3 Internal Dataset: Per-Label Results

The following table preserves the reported validation-set results for 40 labels. AUROC and AUPRC describe discrimination; sensitivity, specificity, and Youden’s J describe class-wise operating points. Performance varies across labels and metrics, with neither pretraining recipe uniformly superior.

Table S1: Consolidated per-label performance on the internal dataset across AUROC, AUPRC, Sensitivity, Specificity, and Youden’s J for Med-CLIP and our autoregressive models (AR-2B, AR-8B). Thresholds for Sensitivity/Specificity/Youden are obtained per abnormality by maximising $J = \text{Sens} + \text{Spec} - 1$. Tail abnormalities are marked with an asterisk (*), and medium-prevalence abnormalities with a dagger (†). For each metric, the best value per row is shown in bold and the second-best is underlined.

Abnormality	AUROC			AUPRC			Sensitivity			Specificity			Youden’s J		
	Med-CLIP	AR-2B (ours)	AR-8B (ours)	Med-CLIP	AR-2B (ours)	AR-8B (ours)	Med-CLIP	AR-2B (ours)	AR-8B (ours)	Med-CLIP	AR-2B (ours)	AR-8B (ours)	Med-CLIP	AR-2B (ours)	AR-8B (ours)
Hydropneumothorax*	0.9899	<u>0.9888</u>	0.9863	0.4445	<u>0.3672</u>	0.3013	0.9641	<u>0.9596</u>	0.9417	<u>0.9634</u>	0.9623	0.9713	0.9275	<u>0.9219</u>	0.9130
Hilar Mass*	0.8907	<u>0.8904</u>	0.8764	0.3714	<u>0.3275</u>	<u>0.3671</u>	0.9107	0.9107	<u>0.8571</u>	<u>0.7424</u>	0.7285	0.7850	0.6531	0.6392	<u>0.6421</u>
Granuloma†	0.9185	0.9481	<u>0.9367</u>	0.0782	0.0863	<u>0.0804</u>	0.8346	<u>0.8508</u>	0.8612	0.8395	0.9091	<u>0.8815</u>	0.6741	0.7599	<u>0.7427</u>
Consolidation	0.8687	0.8902	<u>0.8877</u>	0.5193	0.5746	<u>0.5698</u>	0.7924	<u>0.8128</u>	0.8142	0.7824	0.8091	<u>0.8026</u>	0.5748	0.6219	<u>0.6168</u>
Lesion†	0.5796	0.5431	<u>0.5764</u>	<u>0.0792</u>	0.0744	0.0810	<u>0.6790</u>	0.7433	0.6639	0.4683	0.3474	<u>0.4596</u>	0.1473	0.0907	<u>0.1235</u>
Fibrosis	0.8588	0.8866	<u>0.8821</u>	<u>0.3234</u>	0.3986	<u>0.3849</u>	0.7578	0.7951	<u>0.7865</u>	0.7937	0.8168	<u>0.8131</u>	0.5515	0.6119	<u>0.5996</u>
Scoliosis†	0.8772	0.9162	<u>0.9107</u>	0.2814	0.3580	<u>0.3536</u>	0.7769	0.8056	<u>0.7859</u>	0.8583	<u>0.8867</u>	0.9065	0.6352	<u>0.6923</u>	0.6924
Pleural Plaque*	<u>0.9516</u>	0.9639	0.9475	<u>0.1115</u>	<u>0.0767</u>	0.0679	0.9145	0.9145	<u>0.8547</u>	0.8719	<u>0.9326</u>	0.9475	0.7864	0.8471	<u>0.8022</u>
Clavicle Fracture†	0.9211	<u>0.9065</u>	0.8811	0.4177	<u>0.3338</u>	0.2714	0.8033	<u>0.7912</u>	0.7495	0.9381	<u>0.8933</u>	0.8787	0.7414	<u>0.6845</u>	0.6282
Opacity	0.8576	0.8652	<u>0.8618</u>	0.6665	<u>0.6872</u>	0.6899	0.7772	0.7495	<u>0.7762</u>	0.7759	0.8218	<u>0.7920</u>	0.5531	0.5713	<u>0.5682</u>
Cardiomegaly	0.9336	0.9319	<u>0.9322</u>	0.4962	0.4758	<u>0.4906</u>	0.8997	0.8800	<u>0.8916</u>	0.8219	0.8414	<u>0.8314</u>	<u>0.7216</u>	0.7214	0.7230
Rib Fracture†	<u>0.9215</u>	0.9252	0.9141	<u>0.2046</u>	0.2226	0.1909	0.8190	0.8253	0.8121	<u>0.8672</u>	0.8708	0.8561	<u>0.6862</u>	0.6961	0.6682
Pneumothorax†	0.9708	0.9736	<u>0.9722</u>	<u>0.5327</u>	0.5774	0.5313	0.8738	<u>0.8868</u>	0.8991	<u>0.9503</u>	0.9586	0.9377	0.8241	0.8454	<u>0.8368</u>
Mediastinal Mass*	0.9724	0.9891	<u>0.9750</u>	0.2318	0.1814	<u>0.2082</u>	0.8800	<u>0.9400</u>	0.9467	0.9811	0.9544	<u>0.9554</u>	0.8611	<u>0.8944</u>	0.9021
Cyst*	0.8566	<u>0.8868</u>	0.9137	0.0468	<u>0.0494</u>	0.0550	<u>0.7976</u>	0.7614	0.8265	0.7716	0.8625	<u>0.8607</u>	0.5692	<u>0.6239</u>	0.6872
Ground Glass*	0.8830	<u>0.8683</u>	0.8614	0.0261	0.0165	<u>0.0227</u>	<u>0.8051</u>	0.8159	0.7906	0.8160	<u>0.7800</u>	0.7558	0.6211	<u>0.5959</u>	0.5464
Reticulonodular Pattern†	0.8417	0.8711	<u>0.8588</u>	<u>0.1314</u>	0.1334	0.1246	0.6960	<u>0.7469</u>	0.7654	0.8363	<u>0.8283</u>	0.8033	0.5323	0.5752	<u>0.5687</u>
Pneumonia	0.7784	0.7991	<u>0.7944</u>	0.1157	0.1309	<u>0.1254</u>	0.6299	<u>0.6847</u>	0.7104	0.8039	<u>0.7587</u>	0.7199	<u>0.4338</u>	0.4434	0.4303
Nodule	0.8910	0.9110	<u>0.9066</u>	0.3490	0.4264	<u>0.4062</u>	<u>0.8143</u>	0.8275	0.8112	0.8154	<u>0.8493</u>	0.8540	0.6297	0.6768	<u>0.6652</u>
Scarring*	0.8923	<u>0.9211</u>	0.9271	0.0587	<u>0.0644</u>	0.0759	0.7771	<u>0.8471</u>	0.8726	0.8643	0.8562	<u>0.8598</u>	0.6414	<u>0.7033</u>	0.7324
Pleural Effusion	0.9369	0.9507	<u>0.9485</u>	0.5978	0.6813	<u>0.6728</u>	0.8838	0.9034	<u>0.8916</u>	0.8477	<u>0.8626</u>	0.8682	0.7315	0.7660	<u>0.7598</u>
Mass*	<u>0.9879</u>	0.9922	0.9852	0.2868	<u>0.2856</u>	0.2849	<u>0.9532</u>	0.9640	0.9353	0.9463	<u>0.9531</u>	0.9600	<u>0.8995</u>	0.9171	0.8953
Mediastinal Shift†	0.9447	0.9667	<u>0.9563</u>	<u>0.2132</u>	0.2128	0.2166	0.8424	<u>0.8923</u>	0.9014	<u>0.9193</u>	0.9252	0.8878	0.7617	0.8175	<u>0.7892</u>
Pulmonary Edema†	0.8918	0.9070	<u>0.8946</u>	0.0919	0.0601	<u>0.0681</u>	0.7873	<u>0.8171</u>	0.8260	<u>0.8336</u>	0.8343	0.8086	0.6209	0.6514	<u>0.6346</u>
Hilar Prominence	0.7663	<u>0.7627</u>	0.7535	0.1057	<u>0.0946</u>	0.0940	<u>0.6712</u>	0.6798	0.6631	0.7168	0.6802	<u>0.6966</u>	0.3880	<u>0.3600</u>	0.3597
Bronchitis	0.7573	0.7724	<u>0.7653</u>	0.2653	0.2698	<u>0.2657</u>	0.6225	0.7048	<u>0.6613</u>	0.7561	0.6876	<u>0.7264</u>	0.3786	0.3924	<u>0.3877</u>
Emphysema†	0.9025	0.9136	<u>0.9065</u>	0.2231	0.2101	<u>0.2216</u>	0.7656	0.8243	<u>0.8229</u>	0.8778	<u>0.8531</u>	0.8360	0.6434	0.6774	<u>0.6589</u>
Calcification	0.9188	0.9337	<u>0.9288</u>	0.3730	0.4181	<u>0.4097</u>	0.8274	0.8464	<u>0.8445</u>	0.8695	0.8780	<u>0.8712</u>	0.6969	0.7244	<u>0.7157</u>
Cavity†	0.9459	0.9579	<u>0.9512</u>	0.1694	0.3671	<u>0.2487</u>	0.8715	0.8772	<u>0.8723</u>	0.9055	0.9224	<u>0.9189</u>	0.7770	0.7996	<u>0.7912</u>
Lymphadenopathy*	0.6685	<u>0.6592</u>	0.5931	0.0128	<u>0.0121</u>	0.0098	<u>0.6667</u>	0.7187	0.6496	0.5854	<u>0.5388</u>	0.4933	<u>0.2521</u>	0.2575	0.1429
Hernia†	<u>0.9672</u>	0.9686	<u>0.9672</u>	<u>0.1406</u>	0.1373	0.1426	<u>0.9067</u>	0.9300	0.9026	<u>0.9233</u>	0.9191	0.9447	0.8300	0.8491	<u>0.8473</u>
Normal	<u>0.8549</u>	0.8548	0.8562	<u>0.8542</u>	0.8524	0.8589	<u>0.7367</u>	0.7442	0.7237	<u>0.7863</u>	0.7716	0.7959	0.5230	0.5158	<u>0.5196</u>
Pleural Thickening	0.9090	0.9128	<u>0.9108</u>	0.1522	0.1751	<u>0.1563</u>	0.8504	0.8201	<u>0.8279</u>	0.8192	0.8528	<u>0.8361</u>	<u>0.6696</u>	0.6729	0.6640
Edema*	0.9338	0.9518	<u>0.9457</u>	0.1266	<u>0.0855</u>	0.0847	<u>0.8669</u>	0.8912	0.8523	0.8763	<u>0.8892</u>	0.9081	0.7432	0.7804	<u>0.7604</u>
Tented Diaphragm*	0.8726	<u>0.8962</u>	0.9080	0.0295	0.0181	<u>0.0275</u>	0.6820	<u>0.7972</u>	0.8479	0.9231	<u>0.8387</u>	0.8272	0.6051	<u>0.6359</u>	0.6751
Osteopenia†	0.9223	<u>0.9352</u>	0.9411	<u>0.0993</u>	0.0925	0.1134	0.8592	0.9224	<u>0.8881</u>	<u>0.8363</u>	0.7910	0.8408	0.6955	<u>0.7134</u>	0.7289
Atelectasis	0.9426	0.9532	<u>0.9491</u>	0.1978	0.3358	<u>0.3290</u>	0.8902	0.9131	<u>0.9018</u>	0.8625	<u>0.8632</u>	0.8643	0.7527	0.7763	<u>0.7661</u>
Mediastinal Widening*	0.9347	0.9390	<u>0.9389</u>	0.0831	<u>0.0955</u>	0.0977	0.8095	0.8701	<u>0.8225</u>	<u>0.9097</u>	0.8622	0.9098	<u>0.7192</u>	0.7323	<u>0.7323</u>
Metastasis*	<u>0.9528</u>	0.9569	0.9205	0.5488	<u>0.5417</u>	0.4333	0.9242	0.9242	<u>0.9091</u>	<u>0.8564</u>	0.8763	0.7824	<u>0.7806</u>	0.8005	0.6915
Tracheal Shift†	0.9441	0.9416	<u>0.9428</u>	0.2555	<u>0.2492</u>	0.2389	0.8523	0.8376	<u>0.8404</u>	0.9213	<u>0.9315</u>	0.9362	<u>0.7736</u>	0.7691	0.7766

S4 CheXpert: Per-Label Results

The following tables report the existing results for 115 labels, in the order AUROC, AUPRC, sensitivity, specificity, and Youden’s J .

S4.1 AUROC

Table S2: Per-label AUROC comparison across baseline, contrastive, and autoregressive models on the CheXpert dataset. Tail abnormalities are marked with an asterisk (*), and medium-prevalence abnormalities are marked with a dagger (†). For each row, the best value is shown in bold and the second-best value is underlined.

Abnormality	EVA-Base	RAD-DINO	ARK	CheXFound	BiomedCLIP	MedKLIP	BioViL-T	Med-CLIP	Med-AR-2B (ours)	Med-AR-8B (ours)
LVAD*	0.9166	0.9202	0.9499	<u>0.9741</u>	0.8624	0.9086	0.6904	0.9581	0.9579	0.9893
Diaphragm Elevation	0.7963	0.7187	0.8422	0.8996	0.7299	0.7362	0.5589	<u>0.8998</u>	0.8808	0.9134
Osteopenia†	0.9038	0.8549	0.8978	0.9006	0.8195	0.8803	0.6719	<u>0.9102</u>	0.9052	0.9130
Parenchymal Opacities†	0.7188	0.7463	0.7707	0.7331	0.6791	0.7871	0.7458	<u>0.7794</u>	0.7440	0.7681
Nodules	0.7665	0.7578	0.8370	0.8083	0.7394	0.7590	0.7210	0.8102	0.7904	<u>0.8347</u>
Pleural Effusion	0.8818	0.8527	<u>0.9083</u>	0.9089	0.8191	0.8817	0.7735	0.9053	0.8921	0.9075
Congestive Heart Failure*	0.7978	0.7765	0.8512	0.8580	0.7945	0.8033	0.5773	0.8646	0.8707	<u>0.8667</u>
Prosthetic Valve*	0.8379	0.8322	0.8643	<u>0.8677</u>	0.7147	0.8424	0.5762	0.8577	0.8327	0.8779
Support Devices	0.8492	0.8302	<u>0.8552</u>	0.8551	0.8184	0.8448	0.8160	0.8535	0.8493	0.8595
Picc Line	0.7111	0.6496	<u>0.9529</u>	0.9516	0.6409	0.6623	0.5440	0.9087	<u>0.9531</u>	0.9597
Pleural Thickening	0.7690	0.7497	0.8239	0.8286	0.7269	0.7693	0.6562	<u>0.8395</u>	0.8117	0.8541
Soft Tissue Swelling†	0.8643	0.7971	0.8858	0.9126	0.7717	0.8488	0.4615	0.8981	0.8813	<u>0.9077</u>
Neoplastic Nodules*	0.8542	0.8063	<u>0.9550</u>	0.9495	0.8143	0.8330	0.6700	0.9485	0.9000	0.9682
Nasogastric Tube Removal*	0.7674	0.7300	0.8037	0.8297	0.7558	0.7099	0.6312	<u>0.8351</u>	0.8107	0.8428
Venous Catheter*	0.7332	0.7269	0.7296	0.7899	0.6614	0.7296	0.6284	0.8028	<u>0.8334</u>	0.8364
Hemothorax*	0.7830	0.7981	<u>0.8600</u>	0.8110	0.6953	0.7512	0.6220	0.8817	0.8264	0.8593
CABG†	0.7911	0.7746	0.8238	0.8310	0.7076	0.7513	0.5187	0.8185	<u>0.8382</u>	0.8391
Right Subclavian Line*	0.7321	0.7359	0.7880	0.7820	0.6879	0.7312	0.6872	<u>0.8387</u>	0.8121	0.9119
Aortic Tortuosity†	0.8860	0.8632	0.9102	0.9229	0.8407	0.8722	0.7381	0.9273	<u>0.9286</u>	0.9424
Solitary Nodules†	0.7977	0.7841	<u>0.8548</u>	0.8276	0.7541	0.7829	0.7919	0.8400	0.8332	0.8679
Feeding Tube Removal*	0.6791	0.5639	0.6685	<u>0.7191</u>	0.6868	0.6166	0.6035	0.6845	0.6169	0.7466
Scoliosis†	0.8366	0.7468	0.8232	<u>0.9253</u>	0.7545	0.7590	0.6134	0.9033	0.8589	0.9415
Pulmonary Edema	0.8501	0.8240	<u>0.8706</u>	0.8682	0.7798	0.8457	0.7574	0.8656	0.8492	0.8708
Rib Fractures	0.7168	0.6575	0.8315	0.8213	0.6639	0.6666	0.5499	0.8265	0.7712	<u>0.8307</u>
Ards*	0.9200	0.9055	<u>0.9273</u>	0.9214	0.8517	0.9148	0.8628	0.9240	0.9180	0.9412
Aspiration	0.7089	0.6604	0.7656	<u>0.7798</u>	0.6318	0.6978	0.5871	0.7653	0.7594	0.7809
Reticular Opacities	0.7730	0.7351	0.8075	0.8053	0.6610	0.7793	0.5444	<u>0.8091</u>	0.7833	0.8150
Kyphosis*	0.8945	0.8348	0.9259	0.9522	0.8384	0.8347	0.7309	0.9409	0.9184	0.9500
Vascular Redistribution Cephalization†	0.8127	0.7749	0.8071	0.8011	0.7393	0.7733	0.5645	<u>0.8197</u>	0.7785	0.8434
Airspace Opacity†	0.7196	0.6994	0.7391	<u>0.7412</u>	0.6701	0.7446	0.6788	0.7319	0.6848	0.7347
Left Subclavian Catheter*	0.6514	0.5566	0.7941	0.8431	0.6635	0.5646	0.5941	0.8248	<u>0.9000</u>	0.9354
Kerley B Lines*	0.8169	0.8084	0.8699	0.8742	0.6992	0.7844	0.5527	<u>0.8879</u>	0.8641	0.8944
Chest Tube Removal†	0.7925	0.7513	0.8175	<u>0.8695</u>	0.7198	0.7345	0.5142	0.8546	0.8145	0.8822
Degenerative Changes†	0.8176	0.7939	0.8305	0.8434	0.7774	0.8068	0.7154	<u>0.8477</u>	0.8436	0.8518
Dialysis Catheter*	0.8084	0.7168	0.8797	0.8912	0.8335	0.6677	0.5829	<u>0.9190</u>	0.8968	0.9592
Hilar Lymphadenopathy*	0.7395	0.7471	0.8465	0.8085	0.6727	0.6885	0.6349	<u>0.8773</u>	0.7914	0.8916
Pigtail Catheter†	0.7543	0.7104	<u>0.9481</u>	0.8481	0.6950	0.7274	0.5755	0.8968	0.9251	0.9786
Epidural Catheter†	0.9258	0.8821	0.9510	0.9674	0.8879	0.8669	0.5601	<u>0.9836</u>	0.9829	0.9908
Mediastinal Drain†	0.9368	0.9215	0.9517	<u>0.9542</u>	0.8753	0.9188	0.7046	0.9482	0.9431	0.9644
Retrocardiac Opacity	0.7578	0.7042	0.7830	<u>0.7843</u>	0.6918	0.7380	0.6490	<u>0.7893</u>	0.7686	0.8019

Continued on next page

Abnormality	EVA-Base	RAD-DINO	ARK	CheXFound	BiomedCLIP	MedKLIP	BioViL-T	Med-CLIP	Med-AR-2B (ours)	Med-AR-8B (ours)
Pneumothorax	0.8477	0.8096	0.9410	0.9145	0.7518	0.8585	0.6546	0.9151	0.9103	<u>0.9366</u>
Prosthetic Aortic Valve*	0.7845	0.8111	0.8559	0.8565	0.7901	0.8092	0.6247	0.8469	<u>0.8572</u>	0.8820
Extubation†	0.8111	0.7902	0.8257	0.8268	0.7756	0.7599	0.7213	0.8608	<u>0.8684</u>	0.8797
Perihilar Opacities	0.7605	0.7402	0.8297	0.8354	0.6927	0.7870	0.6533	<u>0.8424</u>	0.8073	0.8463
Sternotomy Wires	0.9020	0.8885	0.9010	0.9044	0.8191	0.8938	0.5702	<u>0.9050</u>	0.8970	0.9105
Subcutaneous Emphysema†	0.9203	0.8292	0.9436	<u>0.9552</u>	0.7612	0.8992	0.5771	0.9532	0.9493	0.9568
MPA Enlargement†	0.7945	0.7179	0.8028	0.8886	0.6877	0.7504	0.5440	<u>0.8939</u>	0.8487	0.9023
Bronchial Wall Thickening*	<u>0.8241</u>	0.7592	0.8106	0.8211	0.7751	0.7826	0.7107	0.8349	0.7931	0.8118
Chest Wall Deformity†	0.6937	0.6386	0.7491	0.7828	0.6113	0.6535	0.6083	<u>0.7968</u>	0.7399	0.8193
Feeding Tube	0.9345	0.8939	0.9453	0.9474	0.8836	0.9358	0.8271	<u>0.9501</u>	0.9489	0.9540
Tracheostomy Cannula*	0.8992	0.7933	0.9747	0.9772	0.7970	0.8382	0.7369	0.9734	0.9736	<u>0.9771</u>
Sternotomy†	0.8674	0.8578	0.8704	<u>0.8735</u>	0.7868	0.8538	0.4712	0.8725	0.8718	0.8901
Granuloma†	0.7746	0.7485	0.8632	0.8020	0.7422	0.7537	0.7262	<u>0.8884</u>	0.8780	0.9194
Mediastinal Mass*	0.6562	0.7037	0.8533	0.8446	0.6021	0.7054	0.6528	<u>0.8561</u>	0.8506	0.8827
Basilar Opacity	0.7283	0.7103	0.7543	<u>0.7613</u>	0.6992	0.7273	0.6826	0.7560	0.7580	0.7699
Artifacts	0.6522	0.5949	0.6659	0.6870	0.6047	0.6206	0.5205	<u>0.6928</u>	0.6690	0.7007
Cardiomegaly	0.8725	0.8181	<u>0.8913</u>	0.8901	0.7813	0.8194	0.6058	0.8895	0.8752	0.8950
Solitary Pulmonary Nodule†	0.7711	0.7583	0.7943	0.7762	0.7377	0.7750	0.7496	<u>0.7996</u>	0.7978	0.8236
Mediastinal Widening	0.8001	0.6869	0.8382	0.8498	0.7043	0.7069	0.5373	<u>0.8620</u>	0.8368	0.8759
Improved Aeration*	0.5931	0.5553	0.5884	0.6413	<u>0.6301</u>	0.5681	0.5315	0.5652	0.5860	0.6239
Chest Tube	0.9264	0.8918	<u>0.9532</u>	0.9467	0.8278	0.8920	0.6535	0.9437	0.9479	0.9599
Malignant Nodules*	0.8661	0.8115	<u>0.9555</u>	0.9452	0.8211	0.8187	0.6661	0.9324	0.8826	0.9589
Tracheostomy Tube†	0.9555	0.8370	<u>0.9756</u>	<u>0.9759</u>	0.8993	0.9425	0.7401	<u>0.9759</u>	0.9749	0.9801
Cancer	0.7715	0.7473	0.8309	0.8358	0.7147	0.7495	0.6059	<u>0.8370</u>	0.8276	0.8650
Pacemaker†	0.9732	0.9419	0.9726	0.9738	0.9295	0.9368	0.5153	<u>0.9755</u>	0.9748	0.9827
Mediastinal Drains†	0.9493	0.9387	<u>0.9620</u>	0.9570	0.9003	0.9330	0.7331	0.9595	0.9577	0.9674
Atrial Dilation*	0.7600	0.7669	<u>0.7952</u>	0.7698	0.7388	0.5822	0.6098	0.7696	0.7613	0.7988
AICD†	0.9665	0.9618	0.9723	0.9789	0.9545	0.9565	0.6053	<u>0.9801</u>	0.9780	0.9818
Subclavian Line*	0.6966	0.6659	0.7601	0.8333	0.6128	0.6588	0.5865	0.8324	<u>0.8795</u>	0.9429
Emphysema†	0.9094	0.8341	0.9233	<u>0.9300</u>	0.8131	0.8945	0.5888	0.9298	0.9253	0.9446
Interstitial Edema	0.7569	0.7259	<u>0.7921</u>	0.7759	0.6852	0.7563	0.6126	0.7870	0.7546	0.7933
Hydropneumothorax*	0.9014	0.8390	0.9338	0.9077	0.7321	0.8786	0.6357	<u>0.9511</u>	0.9162	0.9530
Diaphragmatic Hernia†	0.8113	0.7611	0.8312	<u>0.8982</u>	0.7567	0.7702	0.5783	0.8661	0.8632	0.9059
Pericardial Effusion*	0.7655	0.7115	<u>0.8452</u>	0.8197	0.7153	0.6875	0.4885	0.8543	0.8313	0.8420
Healed Rib Fracture†	0.7463	0.7193	0.8675	0.8342	0.7207	0.7326	0.6600	0.8810	0.7669	<u>0.8792</u>
Alveolar Edema†	0.8627	0.8465	0.8941	0.8854	0.7921	0.8668	0.7051	0.8908	0.8645	<u>0.8913</u>
Aortic Valve Replacement*	0.8268	0.8159	0.8767	<u>0.8775</u>	0.7777	0.8216	0.6329	0.8692	0.8545	0.8915
Hilar Enlargement†	0.6956	0.6542	0.8153	<u>0.8710</u>	0.6114	0.6671	0.5968	0.8591	0.8317	0.8830
Internal Jugular Line	0.8386	0.7987	0.8828	0.8898	0.7750	0.8598	0.6799	<u>0.8904</u>	0.8837	0.8963
Normal	0.9114	0.9061	0.9225	0.9225	0.8746	0.9059	0.8818	<u>0.9211</u>	0.9172	0.9205
Diffused Nodules†	0.8409	0.8076	0.9506	0.9328	0.8166	0.8269	0.6626	0.9393	0.9256	<u>0.9492</u>
Calcified Nodules*	0.7685	0.7498	<u>0.8400</u>	0.7680	0.6903	0.7797	0.7527	0.7950	0.7914	0.8796
Clavicle Fracture†	0.7520	0.7261	0.7847	0.7645	0.6684	0.6941	0.6106	0.9335	0.7440	<u>0.8008</u>
Nasogastric Tube	0.9335	0.9263	0.9464	0.9480	0.9013	0.9413	0.9001	<u>0.9488</u>	0.9432	0.9560
Central Venous Line	0.7840	0.7451	0.8450	0.8437	0.7236	0.7934	0.5626	0.8373	<u>0.8503</u>	0.8599
Left Subclavian Line†	0.7765	0.7325	0.8341	0.8974	0.7124	0.7367	0.6033	0.8648	<u>0.9304</u>	0.9511
Pneumomediastinum†	0.8005	0.6861	<u>0.8727</u>	0.8218	0.7725	0.7776	0.3785	0.8128	0.8096	0.8873
Interstitial Thickening†	0.7520	0.7397	0.7763	0.7583	0.6825	0.7668	0.5192	<u>0.7738</u>	0.7239	0.7719
Pulmonary Vascular Congestion	0.7103	0.6779	0.7402	0.7268	0.6502	0.7001	0.5697	0.7309	0.6984	<u>0.7392</u>
Swan Ganz Catheter	0.8835	0.8768	0.9522	0.9503	0.9183	0.9163	0.7876	<u>0.9552</u>	0.9516	0.9657
Aeration*	0.6051	0.6036	0.5738	0.5811	0.5481	0.5210	0.4089	0.5802	0.6306	<u>0.6224</u>
Mediastinal Shift*	0.9134	0.8151	0.9191	0.9515	0.8745	0.8350	0.5067	0.9576	0.9278	<u>0.9572</u>
Surgical Hardware*	0.6378	0.6314	0.7024	0.6091	0.5865	0.6159	0.4780	0.7109	<u>0.7502</u>	0.7558
Patchy Opacities*	0.7612	0.7208	0.7640	0.7655	0.6777	0.7832	0.6709	0.8215	0.8000	<u>0.8194</u>

Continued on next page

Abnormality	EVA-Base	RAD-DINO	ARK	CheXFound	BiomedCLIP	MedKLIP	BioViL-T	Med-CLIP	Med-AR-2B (ours)	Med-AR-8B (ours)
Mediastinal Clips [†]	0.8610	0.8498	0.8852	0.8988	0.7953	0.8347	0.4249	<u>0.9009</u>	0.8895	0.9118
Opacity	0.6974	0.6762	0.7278	0.7218	0.6487	0.6814	0.6495	0.7243	0.7067	<u>0.7261</u>
Pulmonary Hypertension*	0.7483	0.6552	0.7356	0.8282	0.6695	0.6649	0.5263	<u>0.8486</u>	0.7865	0.8534
Atelectasis	0.6999	0.6854	0.7406	0.7341	0.6449	0.7019	0.6232	0.7304	0.7111	<u>0.7361</u>
Hyperinflation [†]	0.7043	0.6969	0.7057	0.7292	0.6627	0.6827	0.6253	<u>0.7342</u>	0.6750	0.7362
ILD	0.7413	0.7030	<u>0.7797</u>	0.7758	0.6572	0.7386	0.5627	0.7777	0.7481	0.7869
COPD [†]	0.7727	0.7650	0.8151	0.8076	0.7281	0.8003	0.6325	0.8196	0.8016	<u>0.8169</u>
Opacification*	0.8351	0.7904	0.8739	0.8135	0.8401	0.7490	0.6582	0.8506	0.7979	<u>0.8676</u>
Mediport*	0.8160	0.7758	0.9429	0.9307	0.7324	0.6849	0.5639	0.9564	<u>0.9703</u>	0.9884
Lung Transplant*	0.8811	0.8037	0.9180	0.9320	0.7957	0.8303	0.5093	0.9282	<u>0.9389</u>	0.9472
Surgical Clips	0.6955	0.6398	0.8021	0.7760	0.6624	0.6233	0.5709	0.7975	<u>0.8346</u>	0.8892
Low Lung Volumes	0.8326	0.8131	0.8411	0.8428	0.7836	0.8111	0.6937	0.8443	0.8283	<u>0.8436</u>
Aortic Aneurysm [†]	0.7383	0.6037	0.7917	0.8449	0.6906	0.6250	0.5755	<u>0.8727</u>	0.8700	0.9086
Consolidation	0.7348	0.7214	0.7629	0.7558	0.6981	0.7347	0.6911	0.7517	0.7344	<u>0.7579</u>
Pneumonia	0.7344	0.6828	0.7883	<u>0.7855</u>	0.6433	0.7293	0.5692	0.7727	0.7639	0.7853
Free Subdiaphragmatic Air*	0.7206	0.6754	0.7390	0.8138	0.6698	0.6819	0.6444	<u>0.9159</u>	0.8371	0.9286
Fibrosis	0.8414	0.8053	0.8896	0.8893	0.7710	0.8290	0.7200	<u>0.8911</u>	0.8751	0.9016
Lung Mass [†]	0.8102	0.7449	0.8947	0.8906	0.7255	0.7549	0.6157	<u>0.9007</u>	0.8739	0.9098
Endotracheal Tube	0.9454	0.9282	0.9503	0.9541	0.9259	0.9431	0.9163	<u>0.9542</u>	0.9539	0.9598
Pulmonary Arterial Hypertension Signs*	0.8105	0.7498	0.7922	<u>0.8771</u>	0.7009	0.7058	0.6119	0.8895	0.8120	0.8539
Calcification	0.8107	0.7701	0.8439	0.8555	0.7424	0.7796	0.6371	<u>0.8657</u>	0.8341	0.8872

S4.2 AUPRC

Table S3: Per-label AUPRC comparison across baseline, contrastive, and autoregressive models on the CheXpert dataset. Tail abnormalities are marked with an asterisk (*), and medium-prevalence abnormalities are marked with a dagger (†). For each row, the best value is shown in bold and the second-best value is underlined.

Abnormality	EVA-Base	RAD-DINO	ARK	CheXFound	BiomedCLIP	MedKLIP	BioViL-T	Med-CLIP	Med-AR-2B (ours)	Med-AR-8B (ours)
LVAD*	0.0814	0.0851	0.1213	<u>0.2439</u>	0.0848	0.0591	0.0093	0.1771	0.2239	0.3616
Diaphragm Elevation	0.1911	0.1063	0.2607	0.4024	0.1548	0.1251	0.0426	<u>0.4165</u>	0.4013	0.4314
Osteopenia†	<u>0.1545</u>	0.0869	0.1384	0.1384	0.0884	0.1144	0.0248	0.1508	0.1462	0.1724
Parenchymal Opacities†	0.0386	0.0139	0.0139	0.0206	0.0126	<u>0.0255</u>	0.0229	0.0217	0.0168	0.0209
Nodules	0.1140	0.1039	<u>0.2721</u>	0.2329	0.0906	<u>0.1124</u>	0.0725	0.2306	0.2164	0.2727
Pleural Effusion	0.8558	0.8127	0.8902	<u>0.8892</u>	0.7774	0.8552	0.7057	0.8850	0.8715	0.8891
Congestive Heart Failure*	0.0218	0.0234	0.0236	0.0355	0.0172	0.0180	0.0058	0.0602	0.0362	<u>0.0431</u>
Prosthetic Valve*	0.0159	0.0152	0.0172	0.0188	0.0122	0.0111	0.0044	0.0169	0.0182	0.0155
Support Devices	0.2944	0.2592	0.2946	<u>0.2947</u>	0.2518	0.2889	0.2506	0.2880	0.2936	0.3110
Picc Line	0.1639	0.1377	<u>0.6090</u>	0.5874	0.1464	0.1436	0.0897	0.5075	0.5904	0.6169
Pleural Thickening	0.1147	0.0952	0.1542	0.1747	0.0889	0.0985	0.0456	0.1615	0.1578	<u>0.1695</u>
Soft Tissue Swelling†	0.3035	0.1142	<u>0.3336</u>	0.3315	0.1527	0.2508	0.0155	0.3401	0.2831	0.3211
Neoplastic Nodules*	0.0958	0.0403	0.2792	0.2552	0.0548	0.0301	0.0067	<u>0.2977</u>	0.1945	0.3072
Nasogastric Tube Removal*	0.0271	0.0203	<u>0.0357</u>	0.0304	0.0170	0.0151	0.0071	0.0423	0.0272	0.0328
Venous Catheter*	0.0078	0.0096	0.0121	0.0283	0.0073	0.0081	0.0066	0.0194	0.0194	<u>0.0204</u>
Hemothorax*	0.0237	0.0165	0.0705	0.0682	0.0265	0.0096	0.0052	0.0720	0.1020	<u>0.1018</u>
CABG†	0.0691	0.0610	0.0989	<u>0.1155</u>	0.0625	0.0563	0.0253	0.1038	0.1059	0.1214
Right Subclavian Line*	0.0088	0.0099	0.0122	0.0131	0.0071	0.0101	0.0073	0.0199	<u>0.0536</u>	0.0685
Aortic Tortuosity†	0.0996	0.0687	0.1290	0.1623	0.0634	0.0750	0.0290	<u>0.1696</u>	0.1580	0.2187
Solitary Nodules†	0.0195	0.0174	0.0415	0.0495	0.0208	0.0186	0.0166	0.0417	<u>0.0513</u>	0.0701
Feeding Tube Removal*	0.0077	0.0049	0.0076	0.0090	0.0069	0.0053	0.0048	0.0072	0.0050	<u>0.0078</u>
Scoliosis†	0.1196	0.0464	0.1283	<u>0.3646</u>	0.0689	0.0388	0.0178	0.3252	0.2158	0.4447
Pulmonary Edema	0.7112	0.6635	<u>0.7504</u>	0.7461	0.6067	0.7010	0.5489	0.7427	0.7158	0.7505
Rib Fractures	0.1042	0.0619	<u>0.2604</u>	0.2347	0.0791	0.0631	0.0402	0.2865	0.1842	0.2578
ARDS*	0.0742	0.0468	0.0724	0.0478	0.0277	0.0498	0.0154	<u>0.0816</u>	0.0681	0.1012
Aspiration	0.0699	0.0419	0.0839	<u>0.0962</u>	0.0380	0.0522	0.0293	0.0928	0.0773	0.1059
Reticular Opacities	0.2354	0.1840	0.2980	0.2996	0.1617	0.2506	0.0809	<u>0.3062</u>	0.2677	0.3118
Kyphosis*	0.0755	0.0405	0.0598	0.1064	0.0314	0.0255	0.0107	0.1672	0.1158	<u>0.1658</u>
Vascular Redistribution Cephalization†	0.0425	0.0281	0.0432	0.0515	0.0249	0.0290	0.0082	<u>0.0457</u>	0.0373	0.0435
Airspace Opacity†	0.0303	0.0309	0.0490	0.0499	0.0244	0.0378	0.0259	0.0415	0.0235	<u>0.0491</u>
Left Subclavian Catheter*	0.0101	0.0045	0.0141	<u>0.0391</u>	0.0079	0.0056	0.0043	0.0256	0.0352	0.0412
Kerley B Lines*	0.0367	0.0186	0.0454	<u>0.0723</u>	0.0108	0.0195	0.0034	0.1033	0.0287	0.0666
Chest Tube Removal†	0.0597	0.0375	0.0612	<u>0.0866</u>	0.0244	0.0231	0.0060	0.0831	0.0523	0.1153
Degenerative Changes†	0.0953	0.0836	0.1081	<u>0.1205</u>	0.0843	0.0945	0.0478	0.1194	0.1188	0.1262
Dialysis Catheter*	0.0240	0.0067	0.0408	0.0899	0.0284	0.0061	0.0043	0.0834	<u>0.1176</u>	0.1352
Hilar Lymphadenopathy*	0.0119	0.0143	0.0678	<u>0.1143</u>	0.0087	0.0113	0.0074	0.0617	0.0506	0.1172
Pigtail Catheter†	0.0192	0.0174	<u>0.1676</u>	0.0988	0.0211	0.0143	0.0123	0.1614	0.1593	0.1983
Epidural Catheter†	0.1511	0.1151	0.2646	0.2664	0.1618	0.0623	0.0108	0.3916	0.4249	<u>0.4096</u>
Mediastinal Drain†	0.2022	0.1868	0.2301	<u>0.2504</u>	0.1058	0.2036	0.0274	0.2231	0.2130	0.3026
Retrocardiac Opacity	0.1270	0.0982	0.1464	0.1501	0.0905	0.1121	0.0681	<u>0.1589</u>	0.1406	0.1684
Pneumothorax	0.4286	0.3569	0.7242	0.5663	0.3012	0.4442	0.1329	0.6153	0.6045	<u>0.6964</u>
Prosthetic Aortic Valve*	0.0114	0.0116	0.0208	0.0180	0.0144	0.0124	0.0074	0.0417	0.0307	<u>0.0315</u>
Extubation†	0.0640	0.0458	0.0680	0.0646	0.0362	0.0324	0.0237	<u>0.0761</u>	0.0743	0.0885
Perihilar Opacities	0.0866	0.0755	<u>0.1626</u>	0.1560	0.0583	0.1010	0.0415	0.1617	0.1617	0.1857
Sternotomy Wires	0.3066	0.2770	0.2993	<u>0.3168</u>	0.2202	0.2883	0.0810	0.3119	0.2957	0.3266
Subcutaneous Emphysema†	0.1897	0.0705	0.2326	<u>0.2729</u>	0.0856	0.1534	0.0166	0.2661	0.2471	0.2750

Continued on next page

Abnormality	EVA-Base	RAD-DINO	ARK	CheXFound	BiomedCLIP	MedKLIP	BioViL-T	Med-CLIP	Med-AR-2B (ours)	Med-AR-8B (ours)
MPA Enlargement [†]	0.0431	0.0185	0.0678	0.2743	0.0330	0.0281	0.0087	0.2455	<u>0.2870</u>	0.3013
Bronchial Wall Thickening*	0.0344	0.0280	0.0621	0.0451	0.0213	0.0262	0.0133	<u>0.0533</u>	0.0360	0.0489
Chest Wall Deformity [†]	0.0374	0.0200	0.0431	0.0468	0.0163	0.0218	0.0163	0.0450	0.0653	<u>0.0621</u>
Feeding Tube	0.4000	0.2994	0.4684	0.4868	0.3185	0.4258	0.1932	<u>0.4900</u>	0.4891	0.5157
Tracheostomy Cannula*	0.0386	0.0100	0.1304	0.1068	0.0659	0.0328	0.0074	0.0743	0.1089	<u>0.1111</u>
Sternotomy [†]	0.0504	0.0444	0.0523	<u>0.0534</u>	0.0445	0.0448	0.0117	0.0492	0.0509	0.0660
Granuloma [†]	0.0513	0.0397	0.1542	0.0771	0.0402	0.0460	0.0299	0.2658	<u>0.2784</u>	0.3594
Mediastinal Mass*	0.0107	0.0109	0.0640	<u>0.1148</u>	0.0277	0.0096	0.0098	0.0928	0.1136	0.1231
Basilar Opacity	0.1363	0.1265	0.1547	0.1592	0.1300	0.1388	0.1103	0.1578	<u>0.1663</u>	0.1677
Artifacts	0.1236	0.0859	0.1328	<u>0.1652</u>	0.0952	0.1019	0.0668	0.1539	0.1588	0.1734
Cardiomegaly	0.6509	0.5194	<u>0.6820</u>	0.6802	0.4937	0.5205	0.2464	0.6783	0.6560	0.6889
Solitary Pulmonary Nodule [†]	0.0232	0.0256	<u>0.0540</u>	0.0414	0.0217	0.0293	0.0223	0.0479	0.0511	0.0761
Mediastinal Widening	0.1519	0.0550	0.2140	0.2497	0.0734	0.0655	0.0327	<u>0.2669</u>	0.2529	0.2916
Improved Aeration*	0.0097	0.0088	0.0092	<u>0.0124</u>	<u>0.0124</u>	0.0086	0.0086	0.0082	0.0089	0.0143
Chest Tube	0.5163	0.4454	<u>0.5913</u>	0.5621	0.3419	0.3882	0.1051	0.5720	0.5898	0.6285
Malignant Nodules*	0.0836	0.0254	0.2941	0.2682	0.0524	0.0281	0.0064	<u>0.2978</u>	0.2060	0.3151
Tracheostomy Tube [†]	0.3432	0.1055	0.4711	0.4581	0.2830	0.3519	0.0591	0.4678	<u>0.4768</u>	0.4943
Cancer	0.1184	0.0855	0.2157	0.2118	0.0756	0.0914	0.0372	<u>0.2199</u>	0.2064	0.2671
Pacemaker [†]	0.3781	0.2249	0.4018	<u>0.4577</u>	0.3505	0.2188	0.0211	0.4368	0.4423	0.4977
Mediastinal Drains [†]	0.1196	0.0793	0.1648	0.1329	0.0771	0.0692	0.0183	0.1385	0.1318	<u>0.1639</u>
Atrial Dilation*	0.0190	0.0178	0.0317	0.0360	0.0399	0.0055	0.0055	0.0708	0.0177	<u>0.0614</u>
AICD [†]	0.1833	0.1946	0.2393	0.2624	0.1795	0.1606	0.0164	<u>0.2854</u>	0.2656	0.2932
Subclavian Line*	0.0118	0.0119	0.0144	0.0435	0.0080	0.0138	0.0085	0.0231	<u>0.0455</u>	0.0929
Emphysema [†]	0.2474	0.1001	0.2817	0.2910	0.1197	0.2211	0.0207	<u>0.2952</u>	0.2775	0.3057
Interstitial Edema	0.1930	0.1631	0.2251	0.2186	0.1568	0.1831	0.1004	0.2338	0.1873	<u>0.2288</u>
Hydropneumothorax*	0.0378	0.0273	<u>0.1320</u>	0.0853	0.0163	0.0242	0.0046	0.1074	0.1179	0.1355
Diaphragmatic Hernia [†]	0.0356	0.0219	0.0812	<u>0.1978</u>	0.0221	0.0277	0.0079	0.1153	0.1669	0.2598
Pericardial Effusion*	0.0246	0.0103	0.0524	<u>0.0612</u>	0.0141	0.0101	0.0041	0.0501	0.0745	0.0557
Healed Rib Fracture [†]	0.0432	0.0367	0.2191	0.1962	0.0423	0.0407	0.0249	<u>0.2570</u>	0.1632	0.2768
Alveolar Edema [†]	0.0882	0.0664	0.1057	0.1372	0.0594	0.0825	0.0261	0.1037	0.0830	<u>0.1115</u>
Aortic Valve Replacement*	0.0134	0.0117	<u>0.0381</u>	0.0197	0.0128	0.0143	0.0062	0.0194	0.0225	0.0487
Hilar Enlargement [†]	0.0312	0.0324	0.1483	<u>0.2044</u>	0.0253	0.0284	0.0199	0.1914	0.1589	0.2078
Internal Jugular Line	0.2703	0.1942	0.3303	0.3378	0.1970	0.2911	0.1191	<u>0.3400</u>	0.3139	0.3526
Normal	0.4369	0.4037	0.5004	<u>0.5040</u>	0.3097	0.4158	0.3408	0.4987	0.4671	0.5068
Diffused Nodules [†]	0.0708	0.0425	<u>0.2568</u>	0.1905	0.0918	0.0728	0.0099	0.2635	0.1912	0.2303
Calcified Nodules*	0.0130	0.0131	0.0571	0.0258	0.0250	0.0251	0.0114	0.0658	<u>0.1137</u>	0.1205
Clavicle Fracture [†]	0.0350	0.0227	0.0593	0.0547	0.0205	0.0222	0.0127	0.3349	0.1168	<u>0.1567</u>
Nasogastric Tube	0.4794	0.4292	0.5668	<u>0.5874</u>	0.4207	0.5403	0.3565	0.5842	0.5591	0.6208
Central Venous Line	0.2266	0.1635	<u>0.2802</u>	0.2719	0.1621	0.2319	0.0843	0.2672	0.2693	0.3102
Left Subclavian Line [†]	0.0269	0.0183	0.0291	<u>0.1051</u>	0.0247	0.0208	0.0101	0.0586	0.1024	0.1373
Pneumomediastinum*	0.0699	0.0230	0.0745	0.0775	0.0215	0.0771	0.0024	<u>0.0833</u>	0.0578	0.0872
Interstitial Thickening [†]	0.0662	0.0616	<u>0.0808</u>	0.0814	0.0577	0.0761	0.0209	0.0807	0.0754	0.0770
Pulmonary Vascular Congestion	0.1416	0.1119	<u>0.1547</u>	0.1546	0.1108	0.1199	0.0687	0.1527	0.1526	0.1577
Swan Ganz Catheter	0.1584	0.1464	0.3258	0.3693	0.2587	0.2384	0.0665	0.3738	<u>0.4033</u>	0.4627
Aeration*	0.0057	0.0048	0.0058	<u>0.0066</u>	0.0044	0.0039	0.0028	0.0048	0.0095	0.0063
Mediastinal Shift*	0.1320	0.0305	<u>0.2417</u>	0.2226	0.1524	0.0212	0.0057	0.2794	0.1844	0.2347
Surgical Hardware*	0.0097	0.0071	0.0146	0.0059	0.0057	0.0067	0.0076	0.0098	<u>0.0142</u>	0.0136
Patchy Opacities*	0.0218	0.0194	0.0179	0.0190	0.0191	0.0196	0.0130	0.0308	0.0487	<u>0.0383</u>
Mediastinal Clips [†]	0.0553	0.0569	0.0793	0.0715	0.0480	0.0561	0.0080	<u>0.0875</u>	0.0622	0.1161
Opacity	0.1591	0.1477	0.1958	0.1970	0.1412	0.1504	0.1280	<u>0.1989</u>	0.1961	0.2037
Pulmonary Hypertension*	0.0331	0.0138	0.0296	<u>0.1733</u>	0.0189	0.0315	0.0111	0.1142	0.1154	0.1806
Atelectasis	0.5305	0.5114	0.5909	0.5760	0.4743	0.5329	0.4399	0.5789	0.5585	<u>0.5869</u>
Hyperinflation [†]	0.0547	0.0442	0.0445	0.0722	0.0452	0.0523	0.0245	<u>0.0751</u>	0.0694	0.0765
ILD	0.3895	0.3390	0.4628	0.4632	0.3100	0.4053	0.1957	<u>0.4647</u>	0.4241	0.4800

Continued on next page

Abnormality	EVA-Base	RAD-DINO	ARK	CheXFound	BiomedCLIP	MedKLIP	BioViL-T	Med-CLIP	Med-AR-2B (ours)	Med-AR-8B (ours)
COPD [†]	0.0719	0.0344	0.0564	0.0947	0.0334	0.0515	0.0173	0.0824	0.1027	<u>0.0990</u>
Opacification*	0.0311	0.0170	0.0871	0.0586	0.0248	0.0179	0.0069	<u>0.0751</u>	0.0481	0.0651
Mediport*	0.0193	0.0128	0.0612	0.1155	0.0383	0.0121	0.0071	0.1628	<u>0.1989</u>	0.2028
Lung Transplant*	0.0529	0.0310	0.0612	0.0893	0.0470	0.0209	0.0061	0.1473	0.0766	<u>0.1186</u>
Surgical Clips	0.0649	0.0501	0.1679	0.1220	0.0777	0.0453	0.0367	0.1578	<u>0.1907</u>	0.2443
Low Lung Volumes	0.5618	0.5260	0.5758	0.5740	0.4708	0.5296	0.3202	<u>0.5749</u>	0.5605	0.5746
Aortic Aneurysm [†]	0.0967	0.0184	0.1384	0.2168	0.0608	0.0254	0.0180	0.2240	<u>0.2533</u>	0.2674
Consolidation	0.4609	0.4385	<u>0.5014</u>	0.4946	0.4138	0.4565	0.4068	0.4877	0.4684	0.5020
Pneumonia	0.1883	0.1543	0.2845	<u>0.2805</u>	0.1248	0.2001	0.0913	0.2684	0.2564	0.2780
Free Subdiaphragmatic Air*	0.0203	0.0135	0.0160	0.0650	0.0145	0.0101	0.0077	<u>0.2786</u>	0.1955	0.4584
Fibrosis	0.1383	0.0842	0.2556	<u>0.2903</u>	0.0785	0.1118	0.0485	0.2479	0.2759	0.3226
Lung Mass [†]	0.0727	0.0601	0.3225	0.3321	0.0528	0.0531	0.0241	0.3169	<u>0.3584</u>	0.3962
Endotracheal Tube	0.5897	0.4998	0.6330	0.6418	0.5398	0.5984	0.4741	<u>0.6557</u>	0.6548	0.6932
Pulmonary Arterial Hypertension Signs*	0.0378	0.0152	0.0692	0.2534	0.0148	0.0482	0.0070	0.1756	0.2089	<u>0.2250</u>
Calcification	0.1994	0.1426	0.2664	0.2827	0.1277	0.1686	0.0695	<u>0.3088</u>	0.2992	0.3630

S4.3 Sensitivity

Table S4: Per-label **Sensitivity** comparison across baseline, contrastive, and autoregressive models on the CheXpert dataset. Tail abnormalities are marked with an asterisk (*), and medium-prevalence abnormalities are marked with a dagger (†). For each row, the best value is shown in bold and the second-best value is underlined.

Abnormality	EVA-Base	RAD-DINO	ARK	CheXFound	BiomedCLIP	MedKLIP	BioViL-T	Med-CLIP	Med-AR-2B (ours)	Med-AR-8B (ours)
LVAD*	0.9038	0.9423	1.0000	0.9615	0.9038	0.8846	0.8269	0.9231	0.9231	0.9808
Diaphragm elevation	0.7598	0.6520	0.7917	<u>0.8284</u>	0.5858	0.6324	0.8701	0.8064	0.8039	0.8260
Osteopenia†	0.8364	0.8121	0.8000	<u>0.9091</u>	0.8182	0.8182	0.7818	0.8727	0.8545	0.9394
Parenchymal opacities†	0.9286	<u>0.8929</u>	0.8036	0.6964	0.7143	<u>0.8929</u>	0.7679	<u>0.8929</u>	0.6964	0.6250
Nodules	0.6770	0.8217	0.7855	0.7726	0.7829	0.6977	0.7054	0.7519	0.6693	<u>0.7933</u>
Pleural effusion	0.8198	0.7939	<u>0.8434</u>	0.8633	0.7847	0.7970	0.8062	0.8266	0.8227	0.8292
Congestive heart failure*	0.8108	0.6216	0.7568	0.7568	0.6757	0.9459	0.8108	<u>0.9189</u>	0.8378	0.6757
Prosthetic valve*	0.9091	<u>0.9394</u>	0.9697	0.8788	0.6061	0.9091	0.5152	0.8485	0.9091	0.9091
Support devices	0.9295	0.8751	<u>0.9285</u>	0.8731	0.9043	0.8711	0.8872	0.9265	0.9053	0.9144
PICC line	0.7285	0.5321	<u>0.9685</u>	0.9382	0.5297	0.6776	0.6194	0.8461	0.9612	0.9721
Pleural thickening	0.6824	0.6318	0.7770	0.7365	0.6655	0.7703	0.7162	<u>0.7973</u>	0.7196	0.8176
Soft tissue swelling†	0.7356	0.7241	0.7586	0.8161	0.7529	0.7011	0.0230	<u>0.7931</u>	0.7586	0.7816
Neoplastic nodules*	0.8837	<u>0.9070</u>	<u>0.9070</u>	0.8837	0.7907	0.7209	0.8605	<u>0.9070</u>	0.7442	0.9767
Nasogastric tube removal*	0.7931	0.5690	0.6207	0.8793	0.6897	0.8793	<u>0.8966</u>	0.7759	0.7586	0.9138
Venous catheter*	<u>0.9286</u>	0.9048	0.7381	0.9524	0.5952	0.8095	0.4286	0.8571	0.8333	0.8095
Hemothorax*	0.6452	0.7097	0.8387	0.7419	0.5484	<u>0.7742</u>	0.6452	0.6774	0.8387	<u>0.7742</u>
CABG†	0.7689	0.8048	0.8446	0.9203	0.7689	0.8486	0.8845	0.8884	0.8167	<u>0.9163</u>
Right subclavian line*	0.7500	0.5750	0.8250	0.6500	<u>0.9000</u>	0.9500	0.6500	<u>0.9000</u>	0.5750	0.8250
Aortic tortuosity†	0.8654	0.8269	0.8782	0.8846	0.8846	0.8654	0.8462	0.8846	0.9423	<u>0.8974</u>
Solitary nodules†	0.7627	0.7797	0.7966	0.8814	0.7797	0.8136	0.8814	<u>0.8305</u>	0.8136	0.7458
Feeding tube removal*	0.4324	0.5946	0.5135	0.7838	0.7297	0.7568	<u>0.8649</u>	0.8108	0.9189	0.7297
Scoliosis†	0.7122	0.6619	0.8129	0.9065	0.7338	0.7626	0.6187	<u>0.8921</u>	0.7338	0.8849
Pulmonary edema	0.8246	0.8096	0.8467	0.8034	0.7309	<u>0.8285</u>	0.7813	0.8261	0.7955	0.8170
Rib fractures	0.6857	0.6571	<u>0.7740</u>	0.6494	0.7065	0.6597	0.6545	0.8156	0.6571	0.7091
ARDS*	0.8462	<u>0.8974</u>	0.8718	0.9231	0.7692	0.8718	0.9231	0.8205	<u>0.8974</u>	0.9231
Aspiration	0.5462	0.6891	0.6597	0.7017	0.7773	<u>0.7647</u>	0.6471	0.6050	0.7143	0.6849
Reticular opacities	0.6761	0.7005	0.7404	0.7506	0.5900	0.7237	0.8715	<u>0.7815</u>	0.7237	0.7198
Kyphosis*	0.9048	0.8095	0.9762	<u>0.9524</u>	0.7381	0.7857	0.7381	0.9286	0.8571	0.9286
Vascular redistribution cephalization†	0.7692	0.7051	<u>0.7949</u>	0.7821	0.6410	0.6154	0.7436	0.7692	0.6026	0.8333
Airspace opacity†	0.5630	0.7059	0.6218	0.6387	<u>0.7815</u>	0.7059	0.8319	0.5042	0.6387	0.6218
Left subclavian catheter*	0.8108	0.9730	0.8649	0.8108	0.7568	<u>0.8919</u>	0.8649	<u>0.8919</u>	<u>0.8919</u>	<u>0.8919</u>
Kerley B lines*	0.7241	0.6897	0.8621	0.7931	0.6207	0.7586	0.6897	<u>0.8966</u>	0.8621	0.9310
Chest tube removal†	0.7273	0.7576	0.7879	0.8636	0.5606	0.6818	<u>0.8788</u>	0.7879	0.7576	0.9394
Degenerative changes†	0.8377	0.8075	0.8151	0.8264	0.7208	0.7962	0.6151	<u>0.8566</u>	0.8453	0.8642
Dialysis catheter*	0.6857	<u>0.8571</u>	0.8000	0.7143	0.7143	0.4857	0.8000	<u>0.7714</u>	0.7714	0.8857
Hilar lymphadenopathy*	0.8400	0.6800	0.9200	0.6200	0.6200	<u>0.8600</u>	<u>0.8600</u>	0.8000	0.5600	0.8200
Pigtail catheter†	0.8611	0.6944	0.8889	0.7778	0.7222	<u>0.9583</u>	0.9028	0.8472	0.8194	0.9722
Epidural catheter†	0.9271	0.8646	0.9479	0.9792	0.8229	0.8229	<u>0.9583</u>	0.9792	<u>0.9583</u>	0.9792
Mediastinal drain†	0.8571	0.8631	0.9524	0.8929	0.8988	0.9107	0.8750	0.9107	0.8929	<u>0.9226</u>
Retrocardiac opacity	0.7765	0.8314	0.7330	<u>0.8580</u>	0.6989	0.8371	0.8769	0.7784	0.7765	0.7879
Pneumothorax	0.7542	0.7375	0.8904	0.8228	0.6146	0.7508	0.7442	0.8494	0.8283	<u>0.8704</u>
Prosthetic aortic valve*	<u>0.9111</u>	<u>0.9111</u>	0.9556	0.8222	0.7778	<u>0.9111</u>	0.6667	<u>0.9111</u>	0.8000	0.9556
Extubation†	0.7459	0.9098	0.7869	0.7787	0.8443	0.7623	0.8607	0.7869	0.8525	<u>0.8934</u>
Perihilar opacities	0.5870	0.6196	0.7754	0.7391	0.6739	0.7029	0.8370	<u>0.8007</u>	0.7536	0.7572
Sternotomy wires	0.9681	0.9389	<u>0.9806</u>	0.9778	0.7889	0.9708	0.6861	0.9722	0.9750	0.9944

Continued on next page

Abnormality	EVA-Base	RAD-DINO	ARK	CheXFound	BiomedCLIP	MedKLIP	BioViL-T	Med-CLIP	Med-AR-2B (ours)	Med-AR-8B (ours)
Subcutaneous emphysema [†]	0.8605	0.7364	0.8915	0.8760	0.6279	0.7984	0.3721	0.8837	<u>0.9225</u>	0.9302
MPA enlargement [†]	0.6234	0.8701	0.8182	<u>0.8312</u>	0.6234	0.8052	0.5584	0.7532	0.7532	0.8052
Bronchial wall thickening*	0.7407	0.7593	0.7407	0.7593	0.7037	0.7778	0.7222	0.7963	0.8333	<u>0.8148</u>
Chest wall deformity [†]	0.6991	0.3982	<u>0.7611</u>	0.8053	0.5487	0.5487	0.6549	<u>0.7611</u>	0.4956	0.6903
Feeding tube	0.9492	0.9383	0.9492	<u>0.9520</u>	0.8765	0.9259	0.8999	0.9643	0.9438	0.9492
Tracheostomy cannula*	0.8684	0.9211	1.0000	1.0000	0.6053	0.8421	<u>0.9474</u>	1.0000	1.0000	1.0000
Sternotomy [†]	0.9760	0.9520	0.9760	0.9920	0.7040	0.9360	0.1200	<u>0.9840</u>	0.9760	0.9920
Granuloma [†]	0.6849	0.6781	0.7877	0.7397	0.8014	0.7466	0.7740	0.8767	0.7192	<u>0.8219</u>
Mediastinal mass*	0.5636	0.9091	0.8727	0.8000	0.6909	0.8364	<u>0.8909</u>	0.6727	0.7455	0.7273
Basilar opacity	0.7560	0.7404	<u>0.7986</u>	0.7957	0.6383	0.6809	0.7858	0.7631	0.6979	0.8255
Artifacts	0.5829	0.5829	<u>0.6326</u>	0.6000	0.4512	0.4264	0.8899	0.6186	<u>0.6326</u>	0.5659
Cardiomegaly	0.7946	0.7381	<u>0.8269</u>	0.8011	0.7158	0.7346	0.6553	0.8363	0.8145	0.8239
Solitary pulmonary nodule [†]	0.6824	0.8118	0.7647	0.7529	0.6941	0.7647	0.7176	<u>0.7765</u>	0.7647	<u>0.7765</u>
Mediastinal widening	0.7311	0.7180	0.7115	0.7705	0.6852	0.6918	0.9148	0.7180	0.7049	<u>0.8000</u>
Improved aeration*	0.5676	0.5541	0.5541	0.6892	0.6892	<u>0.8378</u>	0.6081	0.9459	0.7027	0.7973
Chest tube	0.8581	0.8431	<u>0.8990</u>	0.8827	0.7763	0.8226	0.7681	0.8772	0.8977	0.9209
Malignant nodules*	<u>0.9091</u>	0.8409	0.9318	0.8409	0.8864	0.7955	0.8182	0.8636	0.8182	<u>0.9091</u>
Tracheostomy tube [†]	0.9070	0.8605	0.9734	0.9701	0.7774	0.8904	0.7674	0.9568	<u>0.9767</u>	0.9834
Cancer	0.6367	0.6609	0.6782	0.7232	0.7163	0.7232	0.5433	<u>0.7924</u>	0.7163	0.7958
Pacemaker [†]	0.9688	0.9107	0.9732	0.9732	0.8661	0.9152	0.9241	<u>0.9866</u>	0.9464	0.9955
Mediastinal drains [†]	<u>0.9878</u>	0.9390	0.9756	1.0000	0.8537	0.9634	0.8049	<u>0.9878</u>	<u>0.9878</u>	0.9756
Atrial dilation*	0.6190	0.7143	0.7857	0.8571	0.8095	0.6190	0.5952	<u>0.8333</u>	0.8095	0.6905
AICD [†]	0.9762	0.9762	<u>0.9921</u>	1.0000	0.9365	0.9603	0.5952	<u>0.9921</u>	0.9603	<u>0.9921</u>
Subclavian line*	0.6604	0.8113	0.8302	0.8113	0.8868	0.7925	0.6981	0.7547	0.7925	<u>0.8679</u>
Emphysema [†]	0.8176	0.8176	0.8706	0.8706	0.6471	0.8176	0.5706	0.9059	<u>0.8824</u>	0.8765
Interstitial edema	0.7993	0.8017	0.7433	0.7786	0.6910	0.8200	0.8224	0.8187	<u>0.8418</u>	0.8504
Hydropneumothorax*	0.8788	0.8182	0.8485	0.8788	0.7576	<u>0.9091</u>	0.8788	0.9394	0.8485	0.8788
Diaphragmatic hernia [†]	0.7246	0.6087	0.7246	<u>0.8986</u>	0.6812	0.8116	0.5652	0.9130	0.8551	0.7971
Pericardial effusion*	0.7872	0.7447	0.8511	0.7660	0.7660	0.7447	0.8511	0.9149	<u>0.8936</u>	0.6596
Healed rib fracture [†]	0.7037	0.6914	0.7346	<u>0.8148</u>	0.6543	0.6852	0.6543	0.8889	0.7222	0.7284
Alveolar edema [†]	0.7699	0.7345	0.7699	0.8673	0.6549	0.7876	0.8319	0.7876	0.7699	<u>0.8584</u>
Aortic valve replacement*	0.9767	0.9070	0.8372	<u>0.9535</u>	0.7907	0.9070	0.6744	0.9767	0.8837	0.9302
Hilar enlargement [†]	0.6807	0.5422	0.7530	0.7771	<u>0.8193</u>	0.7169	0.7530	0.7892	0.7169	0.8855
Internal jugular line	0.8755	<u>0.9098</u>	0.9047	0.8895	0.7903	0.8793	0.7459	0.8831	0.8907	0.9187
Normal	0.8986	0.8784	<u>0.8953</u>	0.8767	0.8649	<u>0.8953</u>	0.8311	0.8378	0.8547	0.8750
Diffused nodules [†]	0.7667	0.6667	<u>0.8667</u>	0.8500	0.6667	0.7833	0.7000	0.8500	0.8500	0.9167
Calcified nodules*	0.7250	0.7750	<u>0.8750</u>	0.6500	0.6250	0.9500	0.6000	0.5250	0.5750	0.8250
Clavicle fracture [†]	0.4773	0.7045	0.7273	0.7045	0.5682	0.5341	0.7159	0.8636	0.5795	<u>0.7727</u>
Nasogastric tube	0.9427	0.9379	0.9510	0.9379	0.9032	<u>0.9450</u>	0.9259	0.9438	0.9223	<u>0.9450</u>
Central venous line	0.8281	0.8790	0.8597	0.8790	0.8047	0.8886	0.7056	0.8776	0.8680	<u>0.8803</u>
Left subclavian line [†]	0.8533	0.7067	<u>0.8933</u>	0.7600	0.7067	0.6800	0.8133	0.7067	0.9200	0.8533
Pneumomediastinum*	0.7500	0.6250	<u>0.7812</u>	0.7500	0.6875	0.7188	0.9688	0.7188	0.6562	<u>0.7812</u>
Interstitial thickening [†]	0.6601	0.6601	<u>0.7044</u>	0.6305	0.6256	0.6995	0.3153	0.7241	0.5961	0.6650
Pulmonary vascular congestion	<u>0.7268</u>	0.5176	0.6773	0.6789	0.5240	0.6054	0.8722	0.6374	0.5479	0.7077
Swan-Ganz catheter	<u>0.8963</u>	0.8506	0.9212	0.8880	0.8174	0.8548	0.8797	0.8714	0.8797	<u>0.8963</u>
Aeration*	0.6486	0.7027	0.7297	<u>0.8919</u>	0.8649	0.6486	1.0000	0.6757	0.8378	0.6486
Mediastinal shift*	0.9123	0.6667	0.8596	<u>0.8947</u>	0.7895	0.8772	0.5614	0.9123	0.8596	0.8596
Surgical hardware*	0.8409	0.5000	0.9091	0.6591	0.7727	0.5227	0.1136	0.6364	0.7727	<u>0.8636</u>
Patchy opacities*	0.7778	0.6190	<u>0.8095</u>	0.7302	0.6190	0.8571	0.6984	0.7619	<u>0.8095</u>	0.7143
Mediastinal clips [†]	0.9053	0.8737	<u>0.9263</u>	0.9053	0.7789	0.8211	0.9895	0.8316	0.8947	0.9158
Opacity	0.8458	0.8153	<u>0.8300</u>	0.7702	0.7775	0.8017	0.8269	0.7828	0.7062	0.6957
Pulmonary hypertension*	0.6964	0.5179	0.6964	<u>0.7321</u>	0.4821	0.6071	0.2321	0.7679	0.6429	0.6786
Atelectasis	0.7538	0.7956	0.7871	0.7714	0.7683	0.8101	<u>0.8093</u>	0.7386	0.7703	0.7120
Hyperinflation [†]	0.4926	0.5735	0.5662	0.6324	0.4191	0.4338	<u>0.6765</u>	0.6691	0.4706	0.7941

Continued on next page

Abnormality	EVA-Base	RAD-DINO	ARK	CheXFound	BiomedCLIP	MedKLIP	BioViL-T	Med-CLIP	Med-AR-2B (ours)	Med-AR-8B (ours)
ILD	0.6978	0.6029	0.7652	<u>0.7707</u>	0.6105	0.6374	0.6780	0.6341	0.7235	0.7888
COPD [†]	0.7021	0.6489	0.7021	<u>0.7660</u>	0.7340	0.7234	0.5532	0.7553	0.8191	0.7128
Opacification*	<u>0.8056</u>	0.6389	0.7500	0.7500	0.7778	0.6389	0.8611	0.6944	<u>0.8056</u>	<u>0.8056</u>
Mediport*	0.8696	0.8043	0.9130	0.8261	0.6087	0.8478	0.3696	0.9348	<u>0.9565</u>	1.0000
Lung transplant*	0.8475	0.7458	<u>0.9322</u>	0.8983	0.6780	0.8475	0.6441	<u>0.9322</u>	0.8814	0.9492
Surgical clips	0.6688	0.6369	0.7325	0.7261	0.7102	0.7389	0.5223	<u>0.7516</u>	<u>0.7771</u>	0.8662
Low lung volumes	0.7502	0.6973	0.7393	0.7507	0.7364	0.7235	0.7007	<u>0.7598</u>	0.7459	0.7869
Aortic aneurysm [†]	0.7535	0.5211	0.7254	0.8310	0.6338	0.6549	0.5915	0.8732	0.7887	<u>0.8451</u>
Consolidation	0.7867	0.8050	<u>0.8605</u>	0.8512	0.7791	<u>0.8605</u>	0.7601	0.8875	0.8177	0.8367
Pneumonia	0.5973	0.6183	0.7720	0.6877	0.6481	0.6010	0.6444	0.6196	0.6927	<u>0.7076</u>
Free subdiaphragmatic air*	0.7045	0.7500	0.6591	0.7727	0.5682	0.6364	<u>0.8182</u>	0.8864	0.7045	0.7955
Fibrosis	0.8339	0.7934	0.8081	0.7675	0.7380	0.7860	0.6384	0.7749	<u>0.8266</u>	<u>0.8266</u>
Lung mass [†]	0.8000	0.7150	0.7650	0.7800	0.6750	0.7400	0.8300	<u>0.8100</u>	0.8050	0.7750
Endotracheal tube	0.9515	0.9451	0.9570	<u>0.9634</u>	0.9121	0.9542	0.9011	0.9542	0.9460	0.9643
Pulmonary arterial hypertension signs*	0.7872	0.7447	0.7021	0.7447	0.6383	0.8511	0.7872	<u>0.8298</u>	0.6596	0.7447
Calcification	0.6925	0.6789	0.7660	<u>0.8143</u>	0.6228	0.7389	0.6905	0.8414	0.7118	0.7795

S4.4 Specificity

Table S5: Per-label **Specificity** comparison across baseline, contrastive, and autoregressive models on the CheXpert dataset. Tail abnormalities are marked with an asterisk (*), and medium-prevalence abnormalities are marked with a dagger (†). For each row, the best value is shown in bold and the second-best value is underlined.

Abnormality	EVA-Base	RAD-DINO	ARK	CheXFound	BiomedCLIP	MedKLIP	BioViL-T	Med-CLIP	Med-AR-2B (ours)	Med-AR-8B (ours)
LVAD*	0.8065	0.7863	0.8010	<u>0.9075</u>	0.7148	0.8121	0.4776	0.8730	0.8487	0.9339
Diaphragm elevation	0.6971	0.6838	0.7580	0.8405	0.7724	0.7352	0.2505	<u>0.8470</u>	0.8271	0.8646
Osteopenia†	<u>0.8344</u>	0.7647	0.8530	0.7553	0.7418	0.7813	0.5046	0.8146	0.8315	0.7632
Parenchymal opacities†	0.4206	0.5634	0.6590	<u>0.6801</u>	0.5716	0.5945	0.6355	0.5439	0.6716	0.8011
Nodules	0.7293	0.5739	<u>0.7430</u>	0.7044	0.6098	0.6963	0.6669	0.7201	0.7802	0.7214
Pleural effusion	0.7838	0.7552	0.8211	0.8014	0.7058	0.8035	0.6161	0.8244	0.8040	<u>0.8214</u>
Congestive heart failure*	0.6798	0.8383	0.8273	<u>0.8566</u>	0.7916	0.5695	0.3998	0.6965	0.8279	0.8973
Prosthetic valve*	0.7597	0.7262	0.7028	0.7465	<u>0.7734</u>	0.7253	0.6287	0.7924	0.6649	0.7571
Support devices	0.6490	0.6782	0.6661	0.7194	0.6327	<u>0.6948</u>	0.6317	0.6662	0.6807	0.6859
PICC line	0.5796	0.6904	0.8667	<u>0.8760</u>	0.6741	0.5623	0.4518	0.8013	0.8517	0.8774
Pleural thickening	0.7402	0.7549	0.7308	0.7794	0.7030	0.6493	0.5405	0.7405	<u>0.7744</u>	0.7502
Soft tissue swelling†	0.9070	0.7672	<u>0.9634</u>	0.9180	0.6804	0.8983	0.9921	0.9304	0.9348	0.9519
Neoplastic nodules*	0.6885	0.6321	<u>0.9259</u>	0.9378	0.7301	0.7914	0.4662	0.8758	0.9235	0.8724
Nasogastric tube removal*	0.6783	<u>0.8275</u>	0.8460	0.6569	0.7104	0.4541	0.3930	0.7936	0.7488	0.6156
Venous catheter*	0.4767	0.4763	0.6425	0.5309	0.6982	0.6031	0.8085	0.6680	0.7358	<u>0.7646</u>
Hemothorax*	0.8127	0.8055	0.8013	0.7962	<u>0.8324</u>	0.6726	0.6389	0.9489	0.7114	0.8283
CABG†	0.7182	0.6645	0.6577	0.6168	0.5686	0.5678	0.1737	0.6337	<u>0.7151</u>	0.6228
Right subclavian line*	0.6601	0.8152	0.6344	0.8180	0.4388	0.3950	0.6834	0.6948	0.9265	<u>0.9068</u>
Aortic tortuosity†	0.7795	0.7671	0.8202	<u>0.8397</u>	0.6854	0.7539	0.5704	0.8356	0.8014	0.8777
Solitary nodules†	0.7465	0.7243	<u>0.7853</u>	0.6724	0.6935	0.6792	0.6155	0.7817	0.7677	0.8530
Feeding tube removal*	0.8434	0.5565	<u>0.7502</u>	0.5867	0.5559	0.4808	0.3464	0.5548	0.3149	0.7093
Scoliosis†	0.7892	0.6971	0.6932	0.8307	0.6453	0.6458	0.5724	0.7513	<u>0.8364</u>	0.8523
Pulmonary edema	0.7168	0.6798	0.7294	0.7703	0.6973	0.7019	0.6140	0.7464	0.7388	<u>0.7611</u>
Rib fractures	0.6362	0.5897	0.7455	0.8426	0.5378	0.5906	0.4469	0.6780	0.7707	<u>0.8128</u>
ARDS*	0.8446	0.8170	<u>0.8683</u>	0.8169	0.8425	0.8491	0.7314	0.8739	0.8272	0.8412
Aspiration	<u>0.7653</u>	0.5543	0.7452	0.7302	0.4375	0.5446	0.5100	0.8064	0.6720	0.7516
Reticular opacities	0.7253	0.6601	<u>0.7278</u>	0.6957	0.6625	0.6927	0.2024	0.6921	0.7084	0.7664
Kyphosis*	0.7583	0.7269	0.7287	<u>0.8735</u>	0.8058	0.7733	0.6532	0.8515	0.9099	0.8698
Vascular redistribution cephalization†	0.7451	0.7558	0.7383	0.7385	0.7566	<u>0.8089</u>	0.4220	0.7904	0.8584	0.7316
Airspace opacity†	<u>0.7661</u>	0.6309	0.7609	0.7413	0.4947	0.6937	0.4445	0.8446	0.6641	0.7325
Left subclavian catheter*	0.4734	0.1708	0.6217	0.7688	0.5558	0.2416	0.3693	0.6429	<u>0.7816</u>	0.8745
Kerley B lines*	0.8742	0.8324	0.8148	<u>0.8643</u>	0.8156	0.8023	0.4513	0.7841	0.7745	0.7431
Chest tube removal†	0.7876	0.6967	0.7505	0.7128	0.8037	<u>0.7906</u>	0.2375	0.7582	0.7776	0.6446
Degenerative changes†	0.7000	0.6771	0.7302	0.7656	0.7319	0.7076	0.7219	<u>0.7414</u>	0.7325	0.7396
Dialysis catheter*	0.7991	0.5263	0.8202	0.9060	0.8054	0.7835	0.3845	0.9098	<u>0.9185</u>	0.9522
Hilar lymphadenopathy*	0.5909	0.7287	0.6084	<u>0.8517</u>	0.6984	0.5329	0.3956	0.8016	0.8874	0.8115
Pigtail catheter†	0.5490	0.6363	0.9105	0.8178	0.6320	0.4139	0.2523	0.8570	0.9484	<u>0.9227</u>
Epidural catheter†	0.8202	0.7651	0.8200	0.8574	0.8134	0.7800	0.2207	0.9231	<u>0.9465</u>	0.9750
Mediastinal drain†	0.8916	0.8455	0.8580	0.9053	0.7089	0.8141	0.4795	0.8627	0.8667	<u>0.9021</u>
Retrocardiac opacity	0.6092	0.4950	0.6923	0.5779	0.5984	0.5298	0.4080	0.6686	0.6438	<u>0.6835</u>
Pneumothorax	0.7697	0.7544	0.8526	<u>0.8544</u>	0.7625	0.8051	0.4877	0.8297	0.8460	0.8552
Prosthetic aortic valve*	0.5757	0.7263	0.6933	<u>0.7898</u>	0.7438	0.6814	0.5739	0.7157	0.8080	0.7178
Extubation†	0.7426	0.5608	0.7274	0.7249	0.5997	0.6610	0.5167	0.8021	0.7451	<u>0.7457</u>
Perihilar opacities	0.8010	0.7425	0.7525	<u>0.7954</u>	0.6191	0.7351	0.4240	0.7282	0.6982	0.7849
Sternotomy wires	0.7642	0.7754	0.7612	0.7577	0.7445	0.7528	0.4391	<u>0.7685</u>	0.7517	0.7591

Continued on next page

Abnormality	EVA-Base	RAD-DINO	ARK	CheXFound	BiomedCLIP	MedKLIP	BioViL-T	Med-CLIP	Med-AR-2B (ours)	Med-AR-8B (ours)
Subcutaneous emphysema [†]	0.8526	0.7636	0.9098	<u>0.9412</u>	0.8285	0.8836	0.7480	0.9234	0.8616	0.9454
MPA enlargement [†]	0.8327	0.4813	0.6594	0.8117	0.6603	0.5971	0.5937	0.9371	0.8398	<u>0.8601</u>
Bronchial wall thickening*	<u>0.7789</u>	0.6975	0.7970	0.7784	0.7113	0.6877	0.6282	0.7581	0.6342	0.7116
Chest wall deformity [†]	0.6428	0.8089	0.6360	0.6533	0.6564	0.7008	0.5331	0.7077	0.8730	<u>0.8098</u>
Feeding tube	0.8259	0.7486	0.8725	<u>0.8890</u>	0.7633	0.8512	0.6616	0.8726	0.8812	0.8948
Tracheostomy cannula*	0.8286	0.6032	0.9437	<u>0.9405</u>	0.8979	0.7213	0.4341	0.9360	0.9235	0.9395
Sternotomy [†]	0.7158	0.7010	0.7128	0.7053	<u>0.7587</u>	0.7428	0.9030	0.7167	0.6973	0.7082
Granuloma [†]	0.7240	0.6968	0.7822	0.7398	0.5717	0.6808	0.5912	0.7324	<u>0.8631</u>	0.8645
Mediastinal mass*	0.7400	0.4473	0.6864	0.7792	0.5139	0.5470	0.3978	<u>0.9002</u>	0.8525	0.9117
Basilar opacity	0.6003	0.5893	0.5854	0.6036	<u>0.6610</u>	0.6597	0.4999	0.6244	0.6890	0.5918
Artifacts	0.6290	0.5485	0.6167	0.6882	0.7141	0.7366	0.1523	0.6726	0.6202	<u>0.7254</u>
Cardiomegaly	0.7952	0.7428	0.7992	0.8218	0.7093	0.7543	0.5011	0.7832	0.7882	<u>0.8098</u>
Solitary pulmonary nodule [†]	0.7722	0.6357	0.6889	0.6799	0.7115	0.6851	0.7197	0.6946	<u>0.7543</u>	0.7317
Mediastinal widening	0.7205	0.5730	0.8095	0.7921	0.6163	0.6192	0.1523	0.8481	<u>0.8385</u>	0.8183
Improved aeration*	<u>0.6118</u>	0.5779	0.6247	0.5429	0.5402	0.2809	0.4731	0.2171	0.4628	0.4115
Chest tube	0.8512	0.7832	0.9020	0.8909	0.7421	0.8107	0.4706	0.8869	0.8866	<u>0.8988</u>
Malignant nodules*	0.6569	0.7051	0.9060	0.9330	0.6749	0.7304	0.5307	0.9086	0.8253	<u>0.9250</u>
Tracheostomy tube [†]	0.9118	0.6905	0.9474	<u>0.9585</u>	0.8657	0.8932	0.6146	0.9523	0.9584	0.9586
Cancer	0.7815	0.7272	0.8404	<u>0.8191</u>	0.6066	0.6512	0.6350	0.7342	0.8087	0.7824
Pacemaker [†]	<u>0.9221</u>	0.8609	0.9135	0.9183	0.8400	0.8662	0.1672	0.9124	0.9321	0.9186
Mediastinal drains [†]	0.8409	0.8610	<u>0.8840</u>	0.8549	0.8400	0.8295	0.5867	0.8726	0.8534	0.9011
Atrial dilation*	<u>0.8044</u>	0.7142	0.6975	0.5688	0.5898	0.6350	0.6170	0.6445	0.5984	0.8311
AICD [†]	0.9136	0.8763	0.9101	0.9200	0.8936	0.8819	0.5800	0.9314	0.9539	<u>0.9442</u>
Subclavian line*	0.6946	0.4753	0.5698	0.7387	0.3108	0.5005	0.4785	0.8225	<u>0.8685</u>	0.8983
Emphysema [†]	0.8409	0.7178	0.8382	<u>0.8752</u>	0.8377	0.8343	0.5991	0.8362	0.8432	0.8916
Interstitial edema	0.5881	0.5421	0.7062	<u>0.6316</u>	0.5899	0.5669	0.3719	0.6210	0.5465	0.6034
Hydropneumothorax*	0.8063	0.6944	<u>0.9011</u>	0.8719	0.6173	0.7671	0.3999	0.8759	0.8666	0.9313
Diaphragmatic hernia [†]	0.8058	0.8104	<u>0.8465</u>	0.7333	0.8048	0.6374	0.6308	0.7235	0.7388	0.8903
Pericardial effusion*	0.6910	0.6259	0.7084	<u>0.8006</u>	0.5906	0.5966	0.2708	0.6865	0.6126	0.8913
Healed rib fracture [†]	0.6677	0.6598	<u>0.8290</u>	0.7378	0.6833	0.6865	0.6159	0.7091	0.6910	0.8809
Alveolar edema [†]	0.8285	0.8146	<u>0.8758</u>	0.7556	0.8040	0.8070	0.5174	0.8776	0.8219	0.7822
Aortic valve replacement*	0.6105	0.7102	0.8308	0.7318	0.6917	0.6688	0.5911	0.7012	0.7344	<u>0.7573</u>
Hilar enlargement [†]	0.6532	0.6855	0.7521	<u>0.8114</u>	0.3657	0.5532	0.4419	0.8022	0.8300	0.7484
Internal jugular line	0.6663	0.5834	0.7217	<u>0.7678</u>	0.6319	0.7057	0.5470	0.7874	0.7595	0.7649
Normal	0.7937	0.8057	0.8246	0.8404	0.7576	0.7875	0.7914	0.8684	<u>0.8581</u>	0.8355
Diffused nodules [†]	0.8153	0.8184	0.9356	0.9111	0.8814	0.7848	0.5928	0.8964	0.8585	<u>0.9264</u>
Calcified nodules*	0.7297	0.6216	0.6766	0.7949	0.7683	0.4964	0.8102	0.9604	<u>0.9537</u>	0.7925
Clavicle fracture [†]	0.8854	0.6844	0.7350	0.7152	0.7078	0.7908	0.4775	<u>0.8552</u>	0.8156	0.7046
Nasogastric tube	0.8317	0.8299	0.8459	0.8608	0.7922	0.8395	0.7711	<u>0.8622</u>	0.8562	0.8836
Central venous line	0.5982	0.5058	0.6815	0.6685	0.5319	0.5351	0.3999	0.6452	<u>0.6979</u>	0.6982
Left subclavian line [†]	0.6147	0.6811	0.6579	<u>0.8906</u>	0.6082	0.7288	0.4205	0.8576	0.8087	0.9373
Pneumomediastinum*	0.7765	0.7100	0.8040	0.8287	0.7643	0.7372	0.0531	0.7984	0.8917	<u>0.8815</u>
Interstitial thickening [†]	0.7194	0.7007	0.7057	0.7428	0.6452	0.7137	0.7344	0.6840	<u>0.7403</u>	0.7368
Pulmonary vascular congestion	0.5896	<u>0.7264</u>	0.6893	0.6661	0.7064	0.6850	0.2695	0.7017	0.7505	0.6551
Swan-Ganz catheter	0.7351	0.7623	0.9009	0.8984	0.8793	0.8314	0.5826	0.9354	<u>0.9412</u>	0.9616
Aeration*	<u>0.5582</u>	0.5495	0.4449	0.2714	0.2480	0.4667	0.0363	0.4840	0.4088	0.5875
Mediastinal shift*	0.7997	0.8492	0.8742	<u>0.9279</u>	0.8394	0.6985	0.5675	0.9083	0.9043	0.9490
Surgical hardware*	0.4012	<u>0.7686</u>	0.4675	0.6127	0.3991	0.7519	0.9651	0.6801	0.6301	0.6247
Patchy opacities*	0.6689	0.7197	0.6281	0.7479	0.6630	0.6015	0.5978	<u>0.7616</u>	0.6786	0.7924
Mediastinal clips [†]	0.6898	0.6860	0.7388	0.7641	0.7113	0.7590	0.0448	0.8458	0.7471	<u>0.7659</u>
Opacity	0.4543	0.4634	0.5143	0.5630	0.4454	0.4786	0.4239	0.5425	<u>0.5980</u>	0.6371
Pulmonary hypertension*	0.6980	0.7732	0.7195	0.8644	0.7788	0.6473	0.8667	0.8589	<u>0.8854</u>	0.9276
Atelectasis	0.5463	0.4786	0.5682	0.5831	0.4513	0.4945	0.4035	<u>0.6055</u>	0.5440	0.6345
Hyperinflation [†]	0.8350	0.7258	0.7801	0.7453	<u>0.8513</u>	0.8733	0.5303	0.6693	0.8446	0.5702

Continued on next page

Abnormality	EVA-Base	RAD-DINO	ARK	CheXFound	BiomedCLIP	MedKLIP	BioViL-T	Med-CLIP	Med-AR-2B (ours)	Med-AR-8B (ours)
ILD	0.6579	0.6939	0.6456	0.6320	0.6185	<u>0.7132</u>	0.4150	0.7659	0.6394	0.6329
COPD [†]	0.7585	0.8004	<u>0.8197</u>	0.7621	0.7141	0.7224	0.6692	0.7817	0.6714	0.8422
Opacification*	0.8094	0.8604	0.8974	0.7688	0.7674	0.7775	0.4240	<u>0.8799</u>	0.6396	0.8291
Mediport*	0.7025	0.6868	0.8857	<u>0.9104</u>	0.8069	0.4599	0.7368	0.8622	0.8901	0.9598
Lung transplant*	0.7771	0.7435	0.7977	<u>0.8615</u>	0.7748	0.7125	0.4123	0.7932	0.9023	0.8539
Surgical clips	0.6295	0.5696	0.7546	0.7034	0.5393	0.4520	0.6012	0.7146	<u>0.7690</u>	0.7884
Low lung volumes	0.7645	0.7774	0.7868	0.7752	0.6811	0.7501	0.5812	<u>0.7775</u>	0.7577	0.7394
Aortic aneurysm [†]	0.6151	0.6450	0.7409	0.7089	0.6658	0.5545	0.5781	0.6964	<u>0.7987</u>	0.8165
Consolidation	0.5772	0.5457	0.5459	0.5432	0.5304	0.5047	0.5379	0.5008	0.5379	<u>0.5521</u>
Pneumonia	0.7436	0.6468	0.6568	<u>0.7499</u>	0.5618	0.7408	0.4655	0.7892	0.6884	0.7143
Free subdiaphragmatic air*	0.6302	0.5687	0.7054	0.7401	0.6842	0.6607	0.4325	0.7931	<u>0.8482</u>	0.9394
Fibrosis	0.7287	0.7280	0.8330	<u>0.8715</u>	0.7100	0.7512	0.7046	0.8751	0.7791	0.8414
Lung mass [†]	0.6917	0.6769	<u>0.8677</u>	0.8537	0.6884	0.6769	0.4062	0.8523	0.7939	0.8994
Endotracheal tube	0.8670	0.8325	0.8879	0.8884	0.8286	0.8566	0.8281	<u>0.8951</u>	0.8848	0.9010
Pulmonary arterial hypertension signs*	0.7279	0.6811	0.8046	0.9328	0.6914	0.4807	0.4189	0.8855	0.8825	<u>0.9045</u>
Calcification	0.7873	0.7345	0.7624	0.7519	0.7473	0.7064	0.5367	0.7361	<u>0.8068</u>	0.8350

S4.5 Youden’s J

Table S6: Per-label **Youden’s J Score** comparison across baseline, contrastive, and autoregressive models on the CheXpert dataset. Tail abnormalities are marked with an asterisk (*), and medium-prevalence abnormalities are marked with a dagger (†). For each row, the best value is shown in bold and the second-best value is underlined.

Abnormality	EVA-Base	RAD-DINO	ARK	CheXFound	BiomedCLIP	MedKLIP	BioViL-T	Med-CLIP	Med-AR-2B (ours)	Med-AR-8B (ours)
LVAD*	0.7103	0.7286	0.8010	<u>0.8690</u>	0.6186	0.6967	0.3045	0.7961	0.7718	0.9147
Diaphragm elevation	0.4569	0.3358	0.5497	<u>0.6689</u>	0.3582	0.3676	0.1206	0.6534	0.6310	0.6906
Osteopenia†	0.6708	0.5768	0.6530	0.6644	0.5600	0.5995	0.2864	<u>0.6873</u>	0.6860	0.7026
Parenchymal opacities†	0.3492	0.4563	<u>0.4626</u>	0.3765	0.2859	0.4874	0.4034	0.4368	0.3680	0.4261
Nodules	0.4063	0.3956	0.5285	0.4770	0.3927	0.3940	0.3723	0.4720	0.4495	<u>0.5147</u>
Pleural effusion	0.6036	0.5491	<u>0.6645</u>	0.6647	0.4905	0.6005	0.4223	0.6510	0.6267	0.6506
Congestive heart failure*	0.4906	0.4599	0.5841	0.6134	0.4673	0.5154	0.2106	<u>0.6154</u>	0.6657	0.5730
Prosthetic valve*	<u>0.6688</u>	0.6656	0.6725	0.6253	0.3795	0.6344	0.1439	0.6409	0.5740	0.6662
Support devices	0.5785	0.5533	<u>0.5946</u>	0.5925	0.5370	0.5659	0.5189	0.5927	0.5860	0.6003
PICC line	0.3081	0.2225	<u>0.8352</u>	0.8142	0.2038	0.2399	0.0712	0.6474	0.8129	0.8495
Pleural thickening	0.4226	0.3867	0.5078	0.5159	0.3685	0.4196	0.2567	<u>0.5378</u>	0.4940	0.5678
Soft tissue swelling†	0.6426	0.4913	0.7220	0.7341	0.4333	0.5994	0.0151	0.7235	0.6934	<u>0.7335</u>
Neoplastic nodules*	0.5722	0.5391	<u>0.8329</u>	0.8215	0.5208	0.5123	0.3267	0.7828	0.6677	0.8491
Nasogastric tube removal*	0.4714	0.3965	0.4667	<u>0.5362</u>	0.4001	0.3334	0.2896	0.5695	0.5074	0.5294
Venous catheter*	0.4053	0.3811	0.3806	0.4833	0.2934	0.4126	0.2371	0.5251	<u>0.5691</u>	0.5741
Hemothorax*	0.4579	0.5152	0.6400	0.5381	0.3808	0.4468	0.2841	<u>0.6263</u>	0.5501	0.6025
CABG†	0.4871	0.4693	0.5023	<u>0.5371</u>	0.3375	0.4164	0.0582	0.5221	0.5318	0.5391
Right subclavian line*	0.4101	0.3902	0.4594	0.4680	0.3388	0.3450	0.3334	<u>0.5948</u>	0.5015	0.7318
Aortic tortuosity†	0.6449	0.5940	0.6984	0.7243	0.5700	0.6193	0.4166	0.7202	<u>0.7437</u>	0.7751
Solitary nodules†	0.5092	0.5040	0.5819	0.5538	0.4732	0.4928	0.4969	0.6122	0.5813	<u>0.5988</u>
Feeding tube removal*	0.2758	0.1511	0.2637	<u>0.3705</u>	0.2856	0.2376	0.2113	0.3656	0.2338	0.4390
Scoliosis†	0.5014	0.3590	0.5061	0.7372	0.3791	0.4084	0.1911	<u>0.6434</u>	0.5702	0.7372
Pulmonary edema	0.5414	0.4894	<u>0.5761</u>	0.5737	0.4282	0.5304	0.3953	0.5725	0.5343	0.5781
Rib fractures	0.3219	0.2468	<u>0.5195</u>	0.4920	0.2443	0.2503	0.1014	0.4936	0.4278	0.5219
ARDS*	0.6908	0.7144	<u>0.7401</u>	0.7400	0.6117	0.7209	0.6545	0.6944	0.7246	0.7643
Aspiration	0.3115	0.2434	0.4049	<u>0.4319</u>	0.2148	0.3093	0.1571	0.4114	0.3863	0.4365
Reticular opacities	0.4014	0.3606	0.4682	0.4463	0.2525	0.4164	0.0739	<u>0.4736</u>	0.4321	0.4862
Kyphosis*	0.6631	0.5364	0.7049	0.8259	0.5439	0.5590	0.3913	0.7801	0.7670	<u>0.7984</u>
Vascular redistribution cephalization†	0.5143	0.4609	0.5332	0.5206	0.3976	0.4243	0.1656	<u>0.5596</u>	0.4610	0.5649
Airspace opacity†	0.3291	0.3368	<u>0.3827</u>	0.3800	0.2762	0.3996	0.2764	0.3488	0.3028	0.3543
Left subclavian catheter*	0.2842	0.1438	0.4866	0.5796	0.3126	0.1335	0.2342	0.5348	<u>0.6735</u>	0.7664
Kerley B lines*	0.5983	0.5221	<u>0.6769</u>	0.6574	0.4363	0.5609	0.1410	0.6807	0.6366	0.6741
Chest tube removal†	0.5149	0.4543	0.5384	<u>0.5764</u>	0.3643	0.4724	0.1163	0.5461	0.5352	0.5840
Degenerative changes†	0.5377	0.4846	0.5453	0.5920	0.4527	0.5038	0.3370	<u>0.5980</u>	0.5778	0.6038
Dialysis catheter*	0.4848	0.3834	0.6202	0.6203	0.5197	0.2692	0.1845	0.6812	<u>0.6899</u>	0.8379
Hilar lymphadenopathy*	0.4309	0.4087	0.5284	0.4717	0.3184	0.3929	0.2556	<u>0.6016</u>	0.4474	0.6315
Pigtail catheter†	0.4101	0.3307	<u>0.7994</u>	0.5956	0.3542	0.3722	0.1551	0.7042	0.7678	0.8949
Epidural catheter†	0.7473	0.6297	0.7679	0.8366	0.6363	0.6029	0.1790	0.9023	<u>0.9048</u>	0.9542
Mediastinal drain†	0.7487	0.7086	<u>0.8104</u>	0.7982	0.6077	0.7248	0.3545	0.7734	0.7596	0.8247
Retrocardiac opacity	0.3857	0.3264	0.4253	0.4359	0.2973	0.3669	0.2849	<u>0.4470</u>	0.4203	0.4714
Pneumothorax	0.5239	0.4919	0.7430	0.6772	0.3771	0.5559	0.2319	0.6791	0.6743	<u>0.7256</u>
Prosthetic aortic valve*	0.4868	0.6374	<u>0.6489</u>	0.6120	0.5216	0.5925	0.2406	0.6268	0.6080	0.6734
Extubation†	0.4885	0.4706	0.5143	0.5036	0.4440	0.4233	0.3774	0.5890	<u>0.5976</u>	0.6391
Perihilar opacities	0.3880	0.3621	0.5279	<u>0.5345</u>	0.2930	0.4380	0.2610	0.5289	0.4518	0.5421
Sternotomy wires	0.7323	0.7143	<u>0.7418</u>	0.7355	0.5334	0.7236	0.1252	0.7407	0.7267	0.7535
Subcutaneous emphysema†	0.7131	0.5000	0.8013	<u>0.8172</u>	0.4564	0.6820	0.1201	0.8071	0.7841	0.8756

Continued on next page

Abnormality	EVA-Base	RAD-DINO	ARK	CheXFound	BiomedCLIP	MedKLIP	BioViL-T	Med-CLIP	Med-AR-2B (ours)	Med-AR-8B (ours)
MPA enlargement [†]	0.4561	0.3514	0.4776	0.6429	0.2837	0.4023	0.1521	0.6903	0.5930	<u>0.6653</u>
Bronchial wall thickening*	0.5196	0.4568	<u>0.5377</u>	<u>0.5377</u>	0.4150	0.4655	0.3504	0.5544	0.4675	0.5264
Chest wall deformity [†]	0.3419	0.2071	0.3971	0.4586	0.2051	0.2495	0.1880	<u>0.4688</u>	0.3686	0.5001
Feeding tube	0.7751	0.6869	0.8217	<u>0.8410</u>	0.6398	0.7771	0.5615	0.8369	0.8250	0.8440
Tracheostomy cannula*	0.6970	0.5243	0.9437	<u>0.9405</u>	0.5032	0.5634	0.3815	0.9360	0.9235	0.9395
Sternotomy [†]	0.6918	0.6530	0.6888	0.6973	0.4627	0.6788	0.0230	0.7007	0.6733	<u>0.7002</u>
Granuloma [†]	0.4089	0.3749	0.5699	0.4795	0.3731	0.4274	0.3652	<u>0.6091</u>	0.5823	0.6864
Mediastinal mass*	0.3036	0.3564	0.5591	0.5792	0.2048	0.3834	0.2887	0.5729	<u>0.5980</u>	0.6390
Basilar opacity	0.3563	0.3297	0.3840	<u>0.3993</u>	0.2993	0.3406	0.2857	0.3875	0.3869	0.4173
Artifacts	0.2119	0.1314	0.2493	0.2882	0.1653	0.1630	0.0422	<u>0.2912</u>	0.2528	0.2913
Cardiomegaly	0.5898	0.4809	<u>0.6261</u>	0.6229	0.4251	0.4889	0.1564	0.6195	0.6027	0.6337
Solitary pulmonary nodule [†]	0.4546	0.4475	<u>0.4536</u>	0.4328	0.4056	0.4498	0.4373	0.4711	0.5190	<u>0.5082</u>
Mediastinal widening	0.4516	0.2910	0.5210	0.5626	0.3015	0.3110	0.0671	<u>0.5661</u>	0.5434	0.6183
Improved aeration*	0.1794	0.1320	0.1788	0.2321	<u>0.2294</u>	0.1187	0.0812	0.1630	0.1655	0.2088
Chest tube	0.7093	0.6263	<u>0.8010</u>	<u>0.7736</u>	0.5184	0.6333	0.2387	0.7641	0.7843	0.8197
Malignant nodules*	0.5660	0.5460	0.8378	0.7739	0.5613	0.5259	0.3489	0.7722	0.6435	<u>0.8341</u>
Tracheostomy tube [†]	0.8188	0.5510	0.9208	0.9286	0.6431	0.7836	0.3820	0.9091	<u>0.9351</u>	0.9420
Cancer	0.4182	0.3881	0.5186	<u>0.5423</u>	0.3229	0.3744	0.1783	0.5266	0.5250	0.5782
Pacemaker [†]	0.8909	0.7716	0.8867	0.8915	0.7061	0.7814	0.0913	<u>0.8990</u>	0.8785	0.9141
Mediastinal drains [†]	0.8287	0.8000	0.8596	0.8549	0.6937	0.7929	0.3916	<u>0.8604</u>	0.8412	0.8767
Atrial dilation*	0.4234	0.4285	<u>0.4832</u>	0.4259	0.3993	0.2540	0.2122	0.4778	0.4079	0.5216
AICD [†]	0.8898	0.8525	0.9022	0.9200	0.8301	0.8422	0.1752	<u>0.9235</u>	0.9142	0.9363
Subclavian line*	0.3550	0.2866	0.4000	0.5500	0.1976	0.2930	0.1766	<u>0.5772</u>	<u>0.6610</u>	0.7662
Emphysema [†]	0.6585	0.5354	0.7088	<u>0.7458</u>	0.4848	0.6519	0.1697	0.7421	0.7256	0.7681
Interstitial edema	0.3874	0.3438	<u>0.4495</u>	0.4102	0.2809	0.3869	0.1943	0.4397	0.3883	0.4538
Hydropneumothorax*	0.6851	0.5126	0.7496	0.7507	0.3749	0.6762	0.2787	0.8153	0.7151	<u>0.8101</u>
Diaphragmatic hernia [†]	0.5304	0.4191	0.5711	0.6319	0.4860	0.4490	0.1960	<u>0.6365</u>	0.5939	0.6874
Pericardial effusion*	0.4782	0.3706	0.5595	<u>0.5666</u>	0.3566	0.3413	0.1219	0.6014	0.5062	0.5509
Healed rib fracture [†]	0.3714	0.3512	0.5636	0.5526	0.3376	0.3717	0.2702	<u>0.5980</u>	0.4132	0.6093
Alveolar edema [†]	0.5984	0.5491	<u>0.6457</u>	0.6229	0.4589	0.5946	0.3493	0.6652	0.5918	0.6406
Aortic valve replacement*	0.5872	0.6172	0.6680	<u>0.6853</u>	0.4824	0.5758	0.2655	0.6779	0.6181	0.6875
Hilar enlargement [†]	0.3339	0.2277	0.5051	0.5885	0.1850	0.2701	0.1949	<u>0.5914</u>	0.5469	0.6339
Internal jugular line	0.5418	0.4932	0.6264	0.6573	0.4222	0.5850	0.2929	<u>0.6705</u>	0.6502	0.6836
Normal	0.6923	0.6841	0.7199	<u>0.7171</u>	0.6225	0.6828	0.6225	0.7062	0.7128	0.7105
Diffused nodules [†]	0.5820	0.4851	<u>0.8023</u>	0.7611	0.5481	0.5681	0.2928	0.7464	0.7085	0.8431
Calcified nodules*	0.4547	0.3966	<u>0.5516</u>	0.4449	0.3933	0.4464	0.4102	0.4854	0.5287	0.6175
Clavicle fracture [†]	0.3627	0.3889	0.4623	0.4197	0.2760	0.3249	0.1934	0.7188	0.3951	<u>0.4773</u>
Nasogastric tube	0.7744	0.7678	0.7969	0.7987	0.6954	0.7845	0.6970	<u>0.8060</u>	0.7785	0.8286
Central venous line	0.4263	0.3848	0.5412	0.5475	0.3366	0.4237	0.1055	0.5228	<u>0.5659</u>	0.5785
Left subclavian line [†]	0.4680	0.3878	0.5512	0.6506	0.3149	0.4088	0.2338	0.5643	<u>0.7287</u>	0.7906
Pneumomediastinum*	0.5265	0.3350	<u>0.5852</u>	0.5787	0.4518	0.4560	0.0219	0.5172	0.5479	0.6627
Interstitial thickening [†]	0.3795	0.3608	<u>0.4101</u>	0.3733	0.2708	0.4132	0.0497	0.4081	0.3364	0.4018
Pulmonary vascular congestion	0.3164	0.2440	0.3666	0.3450	0.2304	0.2904	0.1417	0.3391	0.2984	<u>0.3628</u>
Swan-Ganz catheter	0.6314	0.6129	<u>0.8221</u>	0.7864	0.6967	0.6862	0.4623	0.8068	0.8209	0.8579
Aeration*	0.2068	0.2522	0.1746	0.1633	0.1129	0.1153	0.0363	0.1597	0.2466	0.2361
Mediastinal shift*	0.7120	0.5159	0.7338	0.8226	0.6289	0.5757	0.1289	<u>0.8206</u>	0.7639	0.8086
Surgical hardware*	0.2421	0.2686	0.3766	0.2718	0.1718	0.2746	0.0787	0.3165	<u>0.4028</u>	0.4883
Patchy opacities*	0.4467	0.3387	0.4376	0.4781	0.2820	0.4586	0.2962	0.5235	0.4881	<u>0.5067</u>
Mediastinal clips [†]	0.5951	0.5597	0.6651	0.6694	0.4902	0.5801	0.0343	<u>0.6774</u>	0.6418	0.6817
Opacity	0.3001	0.2787	0.3443	<u>0.3332</u>	0.2229	0.2803	0.2508	0.3253	0.3042	0.3328
Pulmonary hypertension*	0.3944	0.2911	0.4159	0.5965	0.2609	0.2544	0.0988	0.6268	0.5283	<u>0.6062</u>
Atelectasis	0.3001	0.2742	0.3553	<u>0.3545</u>	0.2196	0.3046	0.2128	0.3441	0.3143	0.3465
Hyperinflation [†]	0.3276	0.2993	0.3463	0.3777	0.2704	0.3071	0.2068	0.3384	0.3152	<u>0.3643</u>
ILD	0.3557	0.2968	<u>0.4108</u>	0.4027	0.2290	0.3506	0.0930	0.4000	0.3629	0.4217

Continued on next page

Abnormality	EVA-Base	RAD-DINO	ARK	CheXFound	BiomedCLIP	MedKLIP	BioViL-T	Med-CLIP	Med-AR-2B (ours)	Med-AR-8B (ours)
COPD [†]	0.4606	0.4493	0.5218	0.5281	0.4481	0.4458	0.2224	<u>0.5370</u>	0.4905	0.5550
Opacification*	0.6150	0.4993	0.6474	0.5188	0.5452	0.4164	0.2851	0.5743	0.4452	<u>0.6347</u>
Mediport*	0.5721	0.4911	0.7987	0.7365	0.4156	0.3077	0.1064	0.7970	<u>0.8466</u>	0.9598
Lung transplant*	0.6246	0.4893	0.7299	0.7598	0.4528	0.5600	0.0564	0.7254	<u>0.7837</u>	0.8031
Surgical clips	0.2983	0.2065	0.4871	0.4295	0.2495	0.1909	0.1235	0.4662	<u>0.5461</u>	0.6546
Low lung volumes	0.5147	0.4747	0.5261	0.5259	0.4175	0.4736	0.2819	0.5373	0.5036	<u>0.5263</u>
Aortic aneurysm [†]	0.3686	0.1661	0.4663	0.5399	0.2996	0.2094	0.1696	0.5696	<u>0.5874</u>	0.6616
Consolidation	0.3639	0.3507	0.4064	<u>0.3944</u>	0.3095	0.3652	0.2980	0.3883	0.3556	0.3888
Pneumonia	0.3409	0.2651	<u>0.4288</u>	0.4376	0.2099	0.3418	0.1099	0.4088	0.3811	0.4219
Free subdiaphragmatic air*	0.3347	0.3187	0.3645	0.5128	0.2524	0.2971	0.2507	<u>0.6795</u>	0.5527	0.7349
Fibrosis	0.5626	0.5214	0.6411	0.6390	0.4480	0.5372	0.3430	<u>0.6500</u>	0.6057	0.6680
Lung mass [†]	0.4917	0.3919	0.6327	0.6337	0.3634	0.4169	0.2362	<u>0.6623</u>	0.5989	0.6744
Endotracheal tube	0.8185	0.7776	0.8449	<u>0.8518</u>	0.7407	0.8108	0.7292	0.8493	0.8308	0.8653
Pulmonary arterial hypertension signs*	0.5151	0.4258	0.5067	<u>0.6775</u>	0.3297	0.3318	0.2061	0.7153	0.5421	0.6492
Calcification	0.4798	0.4134	0.5284	0.5662	0.3701	0.4453	0.2272	<u>0.5775</u>	0.5186	0.6145

S5 MIMIC-CXR: Per-Label Results

The evaluated MIMIC-CXR vocabulary contains 110 labels. The available per-label results below cover 108 labels, in the order AUROC, AUPRC, sensitivity, specificity, and Youden’s J ; results for the remaining two labels are not included in these listings.

S5.1 AUROC

Table S7: Per-label AUROC comparison across baseline, contrastive, and autoregressive models on the MIMIC-CXR dataset. Tail abnormalities are marked with an asterisk (*), and medium-prevalence abnormalities are marked with a dagger (†). For each row, the best value is shown in bold and the second-best value is underlined.

Abnormality	EVA-Base	RAD-DINO	ARK	CheXFound	BiomedCLIP	MedKLIP	BioViL-T	Med-CLIP	Med-AR-2B (ours)	Med-AR-8B (ours)
Diaphragm Elevation	0.8174	0.8314	0.8680	0.9072	0.8812	0.6922	0.5298	0.9076	<u>0.9209</u>	0.9251
Parenchymal Opacities†	0.8764	0.8839	0.8993	0.8904	0.8877	0.8674	0.7626	0.8869	0.8893	<u>0.8975</u>
Nodules	0.7210	0.7590	0.8355	0.8155	0.7608	0.7351	0.5837	0.7949	0.8038	<u>0.8303</u>
Pleural Effusion	0.9117	0.9215	0.9339	<u>0.9327</u>	0.9134	0.9150	0.8240	0.9287	0.9303	0.9321
Congestive Heart Failure*	0.8128	0.8403	<u>0.8747</u>	0.8410	0.8649	0.7886	0.6731	0.8774	0.8726	0.8725
Support Devices	0.9173	0.9198	0.9325	0.9317	0.9246	0.9069	0.8886	0.9303	<u>0.9348</u>	0.9371
PICC Line	0.7570	0.8739	<u>0.9749</u>	0.9723	0.9571	0.7100	0.6885	0.9389	0.9703	0.9754
Pleural Thickening	0.7481	0.7853	0.8410	0.8406	0.7738	0.7292	0.6119	0.8408	<u>0.8433</u>	0.8538
Soft Tissue Swelling†	0.8362	0.8617	<u>0.9043</u>	0.8990	0.8810	0.8010	0.5775	0.9027	0.9033	0.9126
Neoplastic Nodules*	0.8326	0.8873	0.9634	0.9488	0.8887	0.8230	0.4787	0.9304	0.9391	<u>0.9594</u>
Hemothorax*	0.8496	0.8937	0.9294	0.9041	0.9219	0.7409	0.6832	0.9149	0.9345	<u>0.9302</u>
CABG†	0.8889	0.9219	0.9157	0.9281	0.9243	0.8536	0.5751	0.9256	<u>0.9306</u>	0.9432
Scarring*	0.7322	0.7442	0.8268	0.8343	0.7480	0.7398	0.4652	0.8227	<u>0.8395</u>	0.8484
Aortic Tortuosity	0.8132	0.8346	0.8635	<u>0.8786</u>	0.8442	0.7932	0.6592	0.8772	0.8784	0.8900
Solitary Nodules†	0.7742	0.7805	0.8414	<u>0.8179</u>	0.7723	0.7890	0.7273	0.8049	0.8140	0.8137
Tracheal Deviation*	0.7055	0.7214	0.7384	0.6920	0.6806	0.6645	0.5510	0.7526	0.7389	<u>0.7424</u>
Scoliosis†	0.8195	0.8073	0.8273	<u>0.9076</u>	0.8473	0.7352	0.6240	0.8858	0.8829	0.9284
Pulmonary Edema	0.8873	0.8901	0.9081	0.9045	0.8909	0.8888	0.8165	0.9005	0.9042	<u>0.9050</u>
Rib Fractures	0.6943	0.7363	<u>0.8838</u>	0.8701	0.7591	0.6720	0.5705	0.8621	0.8606	0.8859
ARDS*	0.9305	0.9402	<u>0.9476</u>	0.9314	0.9372	0.9439	0.8475	0.9472	0.9426	0.9521
Aspiration†	0.7102	0.7391	0.7864	0.7832	0.7507	0.6887	0.6111	0.7804	<u>0.7879</u>	0.7896
Flattened Diaphragm†	0.9090	0.9140	0.9245	0.9218	0.9218	0.8978	0.7031	<u>0.9251</u>	0.9240	0.9351
Reticular Opacities†	0.8402	0.8696	0.9070	0.8881	0.8614	0.8391	0.5574	0.8993	<u>0.9071</u>	0.9105
Kyphosis*	0.8730	0.8812	0.8981	0.8940	0.8923	0.8444	0.4894	0.9004	<u>0.9083</u>	0.9114
Vascular Redistribution Cephalization	0.7786	0.7890	0.8070	0.8020	0.7914	0.7805	0.6883	0.7966	0.7987	<u>0.8058</u>
Degenerative Changes†	0.7789	0.7810	0.7948	<u>0.8149</u>	0.7993	0.7531	0.5496	0.8097	0.8140	0.8232
Dialysis Catheter*	0.7905	0.9112	0.9534	0.9569	0.9218	0.6397	0.4393	0.9420	<u>0.9600</u>	0.9790
Pigtail Catheter†	0.7977	0.8362	<u>0.9631</u>	0.9432	0.9358	0.7714	0.6040	0.9423	0.9584	0.9721
Hilar Lymphadenopathy†	0.7063	0.7685	<u>0.8482</u>	0.8422	0.8114	0.6923	0.5031	0.8284	0.8314	0.8685
Chest Tube Removal*	0.7996	0.8497	0.8972	0.9000	0.8505	0.6606	0.5530	0.9063	0.9254	<u>0.9248</u>
Retrocardiac Opacity†	0.7026	0.7130	<u>0.7747</u>	0.7687	0.7471	0.6648	0.6476	0.7527	0.7441	0.7942
Pneumothorax	0.8648	0.9024	0.9651	0.9474	0.9175	0.8532	0.7191	0.9411	0.9452	<u>0.9555</u>
Vascular Congestion†	0.7411	0.7589	0.7914	0.7641	0.7649	0.7543	0.7114	0.7642	0.7755	<u>0.7882</u>
Esophageal Drainage Tube*	0.8963	0.9067	0.9157	0.9070	0.8919	0.8905	0.7115	0.9016	0.9119	<u>0.9153</u>
Perihilar Opacities	0.7824	0.8014	<u>0.8393</u>	0.8293	0.8201	0.7870	0.6902	0.8259	0.8380	0.8488
Interstitial Lung Disease Pattern	0.7676	0.7848	<u>0.8281</u>	0.8271	0.7907	0.7685	0.6256	0.8205	0.8236	0.8283
Extubation*	0.7936	0.8347	0.8273	0.7982	0.7890	0.7898	0.6817	0.8376	0.8847	<u>0.8790</u>
Sternotomy Wires	0.9589	0.9626	<u>0.9654</u>	0.9650	0.9623	0.9509	0.6693	0.9644	0.9624	0.9662

Continued on next page

Abnormality	EVA-Base	RAD-DINO	ARK	CheXFound	BiomedCLIP	MedKLIP	BioViL-T	Med-CLIP	Med-AR-2B (ours)	Med-AR-8B (ours)
MPA Enlargement [†]	0.7038	0.7326	0.7805	<u>0.8474</u>	0.7650	0.6995	0.6166	0.8237	0.8172	0.8558
Subcutaneous Emphysema [†]	0.8926	0.9405	<u>0.9771</u>	0.9731	0.9483	0.8331	0.6448	0.9599	0.9674	0.9803
Bronchial Wall Thickening*	0.7431	0.7400	<u>0.7800</u>	0.7756	0.7155	0.6695	0.5941	0.7721	0.7727	0.7826
Chest Wall Deformity [†]	0.7617	0.7697	<u>0.8540</u>	0.8560	0.7748	0.7093	0.5840	0.8630	<u>0.8694</u>	0.8858
Feeding Tube	0.9079	0.9134	0.9277	0.9316	0.9333	0.9138	0.8944	0.9241	<u>0.9374</u>	0.9428
Fluid Overload*	0.7961	0.7822	0.8050	0.7746	0.7824	0.7516	0.7185	0.8065	<u>0.8100</u>	0.8104
Sternotomy [†]	0.9556	0.9605	<u>0.9643</u>	0.9636	0.9607	0.9465	0.5883	0.9620	0.9638	0.9671
Granuloma [†]	0.7294	0.7150	0.8355	0.7975	0.7054	0.6857	0.6099	0.8273	<u>0.8457</u>	0.8857
Mediastinal Mass*	0.6958	0.7525	0.8352	0.8261	0.7662	0.6407	0.4884	<u>0.8406</u>	0.8385	0.8699
Artifacts	0.7019	0.6963	0.7347	0.7380	0.7255	0.6549	0.6201	0.7287	<u>0.7492</u>	0.7526
Spinal Hardware*	0.6375	0.6018	<u>0.9361</u>	0.8704	0.7627	0.5862	0.5685	0.8257	0.8979	0.9415
Cardiomegaly	0.8444	0.8376	<u>0.8705</u>	0.8686	0.8556	0.8215	0.7114	0.8667	0.8699	0.8716
Solitary Pulmonary Nodule [†]	0.7780	0.7875	0.8235	0.8103	0.7629	0.7759	0.6806	0.7919	0.8019	<u>0.8182</u>
Empyema*	0.8684	0.9025	0.9247	0.9420	0.8869	0.7805	0.7416	0.9488	<u>0.9504</u>	0.9546
Mediastinal Widening	0.8017	0.7990	0.8691	0.8617	0.8299	0.7572	0.7107	0.8698	<u>0.8750</u>	0.8836
Chest Tube	0.8918	0.9456	<u>0.9748</u>	0.9659	0.9535	0.8629	0.7221	0.9590	0.9715	0.9766
Malignant Nodules*	0.7996	0.8742	0.9547	0.9270	0.8725	0.8042	0.6165	0.9170	0.9267	<u>0.9428</u>
Pacemaker	0.9696	0.9773	0.9801	0.9785	0.9796	0.8988	0.7083	0.9808	0.9822	<u>0.9818</u>
Cancer	0.8119	0.8436	0.8830	0.8844	0.8538	0.7775	0.6163	0.8736	<u>0.8865</u>	0.8955
Tracheostomy Tube [†]	0.8942	0.9475	0.9847	0.9890	0.9720	0.8169	0.7736	0.9873	<u>0.9903</u>	0.9921
Atrial Dilation*	0.7413	0.7190	0.7609	0.7484	0.7221	0.5786	0.4470	0.7372	0.7825	<u>0.7747</u>
MPA Dilation [†]	0.7178	0.7528	0.8000	<u>0.8679</u>	0.7911	0.7220	0.5724	0.8594	0.8455	0.8823
Emphysema [†]	0.8826	0.8989	0.9150	<u>0.9262</u>	0.9007	0.8770	0.7397	0.9223	0.9218	0.9284
Interstitial Edema	0.8334	0.8371	0.8670	0.8586	0.8427	0.8414	0.7330	0.8573	<u>0.8613</u>	0.8596
Diaphragmatic Hernia [†]	0.8046	0.8076	0.8833	<u>0.9248</u>	0.8429	0.7396	0.5169	0.8662	0.9022	0.9296
Cavitary Nodule*	0.8067	0.8881	<u>0.9008</u>	0.8959	0.8426	0.8190	0.6956	0.8906	0.8873	0.9082
Pericardial Effusion [†]	0.8215	0.8216	<u>0.8946</u>	0.8896	0.8681	0.7792	0.6444	0.8936	0.8903	0.9034
Healed Rib Fracture [†]	0.7173	0.7517	0.9171	0.8931	0.7664	0.7299	0.5865	0.9032	0.8752	<u>0.9166</u>
Alveolar Edema [†]	0.8824	0.8997	<u>0.9104</u>	0.9041	0.9011	0.8991	0.8039	0.9149	0.9040	0.9054
Aortic Valve Replacement*	0.9356	0.9390	<u>0.9566</u>	0.9444	0.9206	0.9403	0.5173	0.9499	0.9484	0.9630
Hilar Enlargement	0.7304	0.7735	0.8305	0.8382	0.8038	0.7395	0.5691	0.8446	<u>0.8482</u>	0.8522
Hiatal Hernia*	0.8251	0.8795	0.8851	0.9465	0.8751	0.8033	0.6499	0.8924	<u>0.9470</u>	0.9702
Bronchovascular Crowding*	0.8453	0.8708	0.8907	0.8694	0.8696	0.8333	0.6384	<u>0.8885</u>	0.8835	0.8847
Internal Jugular Line [†]	0.8619	0.8858	0.9409	0.9546	0.9515	0.8651	0.8201	0.9497	<u>0.9612</u>	0.9619
Normal	0.9152	0.9240	0.9327	0.9341	0.9191	0.9137	0.8795	0.9306	0.9322	<u>0.9328</u>
Diffused Nodules*	0.8137	0.8516	0.9426	0.9166	0.8750	0.8460	0.6252	0.9123	0.9159	<u>0.9382</u>
Calcified Nodules*	0.6006	0.6488	0.8096	0.8045	0.7117	0.5008	0.5407	0.8446	0.8663	<u>0.8518</u>
Bibasilar Opacities [†]	0.7493	0.7758	0.8111	0.7916	0.7710	0.7158	0.6394	<u>0.8157</u>	0.8042	0.8203
Clavicle Fracture*	0.6127	0.6351	0.7335	0.7284	0.7303	0.6418	0.5694	0.8176	0.7346	<u>0.7491</u>
Nasogastric Tube	0.9307	0.9369	0.9477	0.9441	0.9412	0.9324	0.9150	0.9391	<u>0.9530</u>	0.9560
Central Venous Line [†]	0.7702	0.8704	0.9109	0.9152	0.9055	0.7396	0.7029	0.9076	<u>0.9257</u>	0.9310
Left Subclavian Line*	0.8428	0.8999	0.9335	0.9340	0.9345	0.8406	0.5907	0.8950	<u>0.9591</u>	0.9783
Pneumomediastinum*	0.8064	0.8175	0.9076	<u>0.9040</u>	0.8441	0.8014	0.4177	0.8597	0.8838	0.8930
Interstitial Thickening [†]	0.8452	0.8491	0.8810	<u>0.8756</u>	0.8459	0.8386	0.6379	0.8733	0.8752	0.8749
Dobhoff Tube [†]	0.9329	0.9616	0.9775	0.9720	0.9704	0.9491	0.8709	0.9598	<u>0.9825</u>	0.9848
Bronchiectasis*	0.8817	0.9004	<u>0.9100</u>	0.9023	0.8668	0.8798	0.6978	0.8736	0.8991	0.9219
Swan Ganz Catheter [†]	0.9378	0.9607	0.9873	0.9750	0.9728	0.9073	0.7127	0.9778	<u>0.9888</u>	0.9915
Bullae*	0.7336	0.8355	0.8572	<u>0.8731</u>	0.8366	0.7760	0.4163	0.8672	0.8606	0.9181
Adenocarcinoma*	0.7173	0.7907	0.8307	0.8171	0.8139	0.6909	0.4593	0.8549	<u>0.8571</u>	0.8617
Mediastinal Shift [†]	0.8841	0.8904	0.9202	0.9335	0.9202	0.8194	0.6301	0.9402	<u>0.9438</u>	0.9525
Mediastinal Clips*	0.9599	0.9446	0.9669	0.9676	0.9642	0.9415	0.4985	0.9663	<u>0.9679</u>	0.9686
Opacity	0.7268	0.7483	0.7739	0.7656	0.7504	0.7357	0.6760	0.7576	0.7621	<u>0.7718</u>
Atelectasis	0.7907	0.8041	0.8373	0.8278	0.8065	0.7927	0.6897	0.8231	0.8296	<u>0.8339</u>
Hyperinflation	0.8987	0.9023	0.9082	0.9130	0.9020	0.8922	0.7943	0.9122	<u>0.9135</u>	0.9171

Continued on next page

Abnormality	EVA-Base	RAD-DINO	ARK	CheXFound	BiomedCLIP	MedKLIP	BioViL-T	Med-CLIP	Med-AR-2B (ours)	Med-AR-8B (ours)
Diffuse Fibrosis*	0.9654	0.9712	0.9875	0.9879	0.9897	0.9672	0.6987	0.9929	0.9896	<u>0.9921</u>
Surgical Clips	0.6927	0.6827	0.8467	0.7927	0.8089	0.6706	0.6009	0.7727	<u>0.8765</u>	0.9166
Low Lung Volumes	0.8456	0.8471	0.8610	<u>0.8646</u>	0.8619	0.8211	0.7091	0.8619	0.8644	0.8676
Aortic Aneurysm*	0.5884	0.7299	0.8056	0.8483	0.7688	0.5658	0.4683	0.8076	0.8244	<u>0.8460</u>
Pneumonia	0.7498	0.7913	0.8314	0.8238	0.7945	0.7725	0.6492	0.8156	0.8232	<u>0.8289</u>
Consolidation	0.7789	0.8102	0.8457	0.8376	0.8106	0.7913	0.7064	0.8295	0.8330	<u>0.8382</u>
Free Subdiaphragmatic Air*	0.7672	0.8137	0.8534	0.8719	0.8642	0.7187	0.5412	0.8861	<u>0.9005</u>	0.9334
Thoracic Spine Degenerative Changes*	0.8280	0.8035	0.8342	0.8228	0.7926	0.7868	0.6204	0.8178	<u>0.8364</u>	0.8470
Port-a-Cath [†]	0.7986	0.9288	0.9827	0.9778	0.9758	0.6577	0.4950	0.9790	<u>0.9871</u>	0.9885
Fibrosis	0.7999	0.8371	<u>0.8888</u>	0.8842	0.8345	0.7953	0.7067	0.8792	0.8847	0.8919
Lung Mass [†]	0.7902	0.8710	<u>0.9127</u>	<u>0.9127</u>	0.8740	0.7576	0.5593	0.9041	0.9092	0.9149
Endotracheal Tube	0.9637	0.9699	0.9751	0.9750	0.9730	0.9664	0.9499	0.9725	<u>0.9781</u>	0.9794
Pulmonary Arterial Hypertension Signs [†]	0.7149	0.7699	0.7881	0.8282	0.7778	0.7361	0.6329	<u>0.8373</u>	0.8352	0.8538
Effusion	0.9118	0.9213	0.9337	<u>0.9326</u>	0.9127	0.9142	0.8221	0.9282	0.9301	0.9322
Calcification	0.7826	0.7938	0.8460	0.8668	0.8142	0.7568	0.6730	0.8668	<u>0.8687</u>	0.8835
Degenerative Spine Changes [†]	0.8031	0.8204	0.8447	0.8525	0.8355	0.7811	0.5151	0.8405	0.8649	<u>0.8637</u>

S5.2 AUPRC

Table S8: Per-label AUPRC comparison across baseline, contrastive, and autoregressive models on the MIMIC-CXR dataset. Tail abnormalities are marked with an asterisk (*), and medium-prevalence abnormalities are marked with a dagger (†). For each row, the best value is shown in bold and the second-best value is underlined.

Abnormality	EVA-Base	RAD-DINO	ARK	CheXFound	BiomedCLIP	MedKLIP	BioViL-T	Med-CLIP	Med-AR-2B (ours)	Med-AR-8B (ours)
Diaphragm Elevation	0.2024	0.2282	0.3259	0.4529	0.3923	0.0722	0.0334	0.4595	<u>0.4844</u>	0.4967
Parenchymal Opacities†	0.0581	0.0552	0.0691	0.0632	0.0583	0.0502	0.0192	0.0549	0.0630	<u>0.0633</u>
Nodules	0.0991	0.1335	0.3179	0.2356	0.1577	0.1118	0.0422	0.2503	0.2624	<u>0.2965</u>
Pleural Effusion	0.8302	0.8463	0.8706	<u>0.8682</u>	0.8332	0.8343	0.6555	0.8605	0.8663	0.8674
Congestive Heart Failure*	0.0179	0.0190	0.0242	0.0186	0.0197	0.0113	0.0049	<u>0.0241</u>	0.0210	0.0220
Support Devices	0.1770	0.1640	0.1957	<u>0.2010</u>	0.1856	0.1477	0.1401	0.1972	0.1978	0.2053
Picc Line	0.1270	0.2494	0.6183	0.6005	0.5522	0.0928	0.0862	0.4965	0.6083	<u>0.6171</u>
Pleural Thickening	0.0715	0.0999	0.1886	0.2010	0.1159	0.0725	0.0373	0.2072	<u>0.2294</u>	0.2429
Soft Tissue Swelling†	0.2236	0.3382	0.4742	0.4699	0.4079	0.2401	0.0202	0.4624	<u>0.4918</u>	0.5041
Neoplastic Nodules*	0.0440	0.0737	0.4225	0.3393	0.1939	0.0999	0.0040	0.3755	0.3528	<u>0.4110</u>
Hemothorax*	0.0359	0.0357	0.0906	0.0704	0.0825	0.0086	0.0065	0.0838	0.1106	<u>0.1075</u>
CABG†	0.0965	0.1074	0.1105	0.1263	0.1185	0.0882	0.0143	0.1271	<u>0.1446</u>	0.1604
Scarring*	0.0080	0.0134	0.0183	<u>0.0242</u>	0.0095	0.0081	0.0028	0.0241	0.0209	0.0247
Aortic Tortuosity	0.1731	0.2006	0.2460	<u>0.2945</u>	0.2301	0.1600	0.0744	0.2798	0.2807	0.2965
Solitary Nodules†	0.0177	0.0200	0.0510	0.0407	0.0222	0.0212	0.0120	0.0442	<u>0.0463</u>	0.0438
Tracheal Deviation*	0.0070	0.0086	0.0260	0.0086	0.0094	0.0069	0.0040	<u>0.0239</u>	0.0228	0.0229
Scoliosis†	0.1075	0.1002	0.1150	<u>0.4239</u>	0.1843	0.0362	0.0198	0.3310	0.3032	0.4370
Pulmonary Edema	0.6478	0.6586	0.7119	<u>0.7053</u>	0.6661	0.6458	0.4713	0.6905	0.7007	0.7026
Rib Fractures	0.0579	0.0754	0.3188	0.2800	0.1034	0.0501	0.0310	0.2826	0.2635	<u>0.3110</u>
ARDS*	0.0800	0.0736	0.0882	0.0973	0.1259	0.0759	0.0328	0.1107	0.0987	<u>0.1148</u>
Aspiration†	0.0426	0.0467	0.0718	0.0671	0.0584	0.0329	0.0239	0.0645	0.0814	<u>0.0785</u>
Flattened Diaphragm†	0.0825	0.1012	0.1015	<u>0.1308</u>	0.1165	0.0765	0.0117	0.1266	0.1236	0.1373
Reticular Opacities†	0.0698	0.0788	0.1119	<u>0.1400</u>	0.1176	0.0699	0.0065	0.1208	0.1262	0.1446
Kyphosis*	0.0735	0.0511	0.0932	0.1243	<u>0.1253</u>	0.0323	0.0036	0.1194	0.1249	0.1538
Vascular Redistribution Cephalization	0.1394	0.1474	0.1608	0.1561	0.1537	0.1414	0.0858	0.1522	<u>0.1575</u>	0.1555
Degenerative Changes†	0.0399	0.0391	0.0436	0.0541	0.0469	0.0329	0.0141	0.0443	0.0486	<u>0.0539</u>
Dialysis Catheter*	0.0113	0.0666	0.2079	0.2458	0.2383	0.0057	0.0031	0.1679	<u>0.2671</u>	0.3351
Pigtail Catheter*	0.0257	0.0351	<u>0.2607</u>	0.1674	0.2067	0.0234	0.0101	0.1878	0.2788	0.2568
Hilar Lymphadenopathy†	0.0244	0.0383	<u>0.1049</u>	0.0985	0.0791	0.0225	0.0089	0.0941	0.0933	0.1192
Chest Tube Removal*	0.0143	0.0242	0.0390	0.0669	0.0545	0.0060	0.0038	0.1120	0.1385	<u>0.1353</u>
Retrocardiac Opacity†	0.0283	0.0189	<u>0.0369</u>	0.0371	0.0248	0.0136	0.0133	0.0272	0.0267	0.0351
Pneumothorax	0.2790	0.3834	0.7375	0.5616	0.4513	0.2739	0.0993	0.5777	0.5707	<u>0.6830</u>
Vascular Congestion†	0.0243	0.0264	0.0286	0.0261	<u>0.0331</u>	0.0242	0.0206	0.0261	0.0381	0.0280
Esophageal Drainage Tube*	0.0188	0.0189	<u>0.0249</u>	0.0199	0.0210	0.0227	0.0085	0.0267	0.0211	0.0231
Perihilar Opacities	0.0786	0.0867	0.1473	0.1374	0.1277	0.0741	0.0458	0.1404	<u>0.1621</u>	0.1707
Interstitial Lung Disease Pattern	0.3166	0.3426	0.4286	<u>0.4287</u>	0.3772	0.3266	0.1487	0.4176	0.4229	0.4301
Extubation*	0.0153	0.0178	0.0145	0.0191	0.0118	0.0114	0.0066	0.0158	<u>0.0264</u>	0.0267
Sternotomy Wires	0.3341	0.3315	0.3415	<u>0.3462</u>	0.3298	0.2880	0.0494	0.3465	0.3259	0.3451
MPA Enlargement†	0.0201	0.0229	0.0535	<u>0.1105</u>	0.0471	0.0178	0.0104	0.0972	0.1020	0.1529
Subcutaneous Emphysema†	0.1084	0.1976	0.2747	0.3390	0.2398	0.1028	0.0137	0.2627	<u>0.3296</u>	0.3153
Bronchial Wall Thickening*	0.0160	0.0176	0.0234	<u>0.0260</u>	0.0235	0.0108	0.0066	0.0231	0.0240	0.0309
Chest Wall Deformity†	0.0347	0.0357	0.0665	<u>0.0871</u>	0.0436	0.0281	0.0107	0.0809	0.0780	0.1031
Feeding Tube	0.1501	0.1602	0.2075	0.2156	0.2364	0.1604	0.1349	0.2123	<u>0.2573</u>	0.2866
Fluid Overload*	0.0187	0.0171	0.0163	0.0156	0.0146	0.0108	0.0099	0.0170	0.0187	<u>0.0182</u>
Sternotomy†	0.2050	0.2170	0.2365	0.2266	0.2191	0.1922	0.0204	<u>0.2429</u>	0.2342	0.2571
Granuloma†	0.0294	0.0223	0.1820	0.1298	0.0270	0.0200	0.0125	0.1870	<u>0.2335</u>	0.2724

Continued on next page

Abnormality	EVA-Base	RAD-DINO	ARK	CheXFound	BiomedCLIP	MedKLIP	BioViL-T	Med-CLIP	Med-AR-2B (ours)	Med-AR-8B (ours)
Mediastinal Mass*	0.0126	0.0199	0.0998	0.1073	0.0596	0.0093	0.0055	<u>0.1212</u>	0.1190	0.1487
Artifacts	0.1160	0.0972	0.1334	<u>0.1487</u>	0.1304	0.0800	0.0665	0.1341	0.1497	0.1476
Spinal Hardware*	0.0085	0.0039	<u>0.1289</u>	0.0907	0.0464	0.0039	0.0032	0.0709	0.1165	0.1727
Cardiomegaly	0.6932	0.6877	<u>0.7447</u>	0.7424	0.7206	0.6526	0.4592	0.7397	<u>0.7453</u>	0.7489
Solitary Pulmonary Nodule [†]	0.0213	0.0241	<u>0.0502</u>	0.0456	0.0250	0.0240	0.0121	0.0394	0.0386	0.0558
Empyema*	0.0173	0.0307	0.0562	0.0581	0.0510	0.0085	0.0072	0.0654	0.0823	<u>0.0801</u>
Mediastinal Widening	0.1222	0.1181	0.1975	0.1884	0.1438	0.0754	0.0564	0.2021	<u>0.2236</u>	0.2368
Chest Tube	0.2596	0.4147	0.5607	0.4717	0.4508	0.2058	0.0555	0.4424	0.5292	<u>0.5425</u>
Malignant Nodules*	0.0440	0.0863	0.3912	0.2806	0.1655	0.0418	0.0077	0.3163	0.3028	<u>0.3769</u>
Pacemaker	0.3571	0.3789	0.4491	0.4153	0.4511	0.1913	0.0498	0.4559	<u>0.4756</u>	0.4842
Cancer	0.1648	0.2177	<u>0.3205</u>	0.3023	0.2554	0.1267	0.0485	0.2873	0.3141	0.3464
Tracheostomy Tube [†]	0.2039	0.3889	0.7064	0.7306	0.6461	0.0795	0.0606	0.6866	<u>0.7437</u>	0.7593
Atrial Dilation*	0.0071	0.0062	0.0118	0.0091	0.0125	0.0044	0.0023	0.0106	0.0154	<u>0.0141</u>
MPA Dilation [†]	0.0169	0.0192	0.0412	0.0891	0.0271	0.0139	0.0068	0.0851	<u>0.0911</u>	0.1704
Emphysema [†]	0.2063	0.2257	0.2655	0.2822	0.2404	0.1691	0.0579	0.2716	<u>0.2903</u>	0.3274
Interstitial Edema	0.1984	0.2047	0.2634	0.2527	0.2311	0.2185	0.1012	0.2515	0.2562	<u>0.2626</u>
Diaphragmatic Hernia [†]	0.0743	0.0876	0.2256	<u>0.4658</u>	0.1330	0.0457	0.0143	0.1711	0.3807	0.5783
Cavitary Nodule*	0.0169	0.0248	0.0367	0.0298	0.0276	0.0197	0.0063	0.0382	<u>0.0461</u>	0.0483
Pericardial Effusion [†]	0.0855	0.0459	0.1639	0.1606	0.1104	0.0250	0.0120	0.1662	<u>0.1823</u>	0.1926
Healed Rib Fracture [†]	0.0440	0.0539	<u>0.3182</u>	0.2653	0.0635	0.0437	0.0195	0.2968	0.2553	0.3444
Alveolar Edema [†]	0.0540	0.0772	0.0942	0.0774	0.0826	0.0785	0.0215	0.0904	<u>0.0952</u>	0.1089
Aortic Valve Replacement*	0.0428	0.0396	<u>0.0799</u>	0.0405	0.0428	0.0423	0.0039	0.0546	0.0500	0.0882
Hilar Enlargement	0.0709	0.0880	0.1893	0.2125	0.1402	0.0686	0.0283	0.2023	<u>0.2211</u>	0.2277
Hiatal Hernia*	0.0142	0.0204	0.0685	0.1759	0.0607	0.0108	0.0041	0.0413	<u>0.1997</u>	0.2812
Bronchovascular Crowding*	0.0305	0.0406	0.0581	0.0475	<u>0.0525</u>	0.0334	0.0095	0.0390	0.0452	0.0465
Internal Jugular Line [†]	0.0811	0.0907	0.1694	0.2193	0.2237	0.0909	0.0765	0.2034	0.2518	<u>0.2449</u>
Normal	0.7137	0.7445	0.7720	0.7757	0.7237	0.7137	0.6128	0.7662	0.7678	<u>0.7723</u>
Diffused Nodules*	0.0437	0.0896	0.4378	0.2301	0.1839	0.0996	0.0082	0.3240	0.3656	<u>0.4089</u>
Calcified Nodules*	0.0037	0.0043	0.0559	0.0500	0.0125	0.0025	0.0028	0.1021	<u>0.1101</u>	0.1176
Bibasilar Opacities [†]	0.0190	0.0211	<u>0.0385</u>	0.0274	0.0291	0.0149	0.0103	0.0492	0.0319	0.0374
Clavicle Fracture*	0.0077	0.0101	0.0187	0.0217	0.0246	0.0092	0.0068	0.1386	0.0198	<u>0.0441</u>
Nasogastric Tube	0.3462	0.3706	0.4428	0.4220	0.4190	0.3627	0.3023	0.3778	<u>0.4899</u>	0.5046
Central Venous Line [†]	0.0490	0.0926	0.1664	0.1507	0.1576	0.0478	0.0464	0.1627	<u>0.1944</u>	0.2154
Left Subclavian Line*	0.0157	0.0361	0.0432	0.0862	0.0987	0.0137	0.0048	0.0293	<u>0.1395</u>	0.1675
Pneumomediastinum*	0.0656	0.0696	<u>0.1993</u>	0.1967	0.1439	0.0896	0.0038	0.1281	0.1691	0.2008
Interstitial Thickening [†]	0.1221	0.1276	0.1771	0.1609	0.1563	0.1367	0.0238	0.1646	<u>0.1648</u>	0.1624
Dobhoff Tube [†]	0.1245	0.1956	0.3380	0.2584	0.2561	0.1279	0.0359	0.1803	0.3751	<u>0.3676</u>
Bronchiectasis*	0.0413	0.0586	0.0914	0.0659	0.0492	0.0332	0.0061	0.0681	0.1246	<u>0.1049</u>
Swan Ganz Catheter [†]	0.1580	0.2207	0.5597	0.5546	0.6003	0.1058	0.0228	0.5337	<u>0.6929</u>	0.7261
Bullae*	0.0123	0.0247	0.0360	<u>0.0588</u>	0.0183	0.0091	0.0023	0.0449	0.0353	0.0829
Adenocarcinoma*	0.0085	0.0102	0.0102	0.0106	0.0089	0.0055	0.0024	0.0132	0.0184	<u>0.0169</u>
Mediastinal Shift [†]	0.1671	0.0893	0.2091	0.2400	0.2283	0.0352	0.0137	<u>0.2888</u>	0.2692	0.3050
Mediastinal Clips*	0.1378	0.0947	<u>0.1650</u>	0.1451	0.1501	0.0967	0.0051	0.1405	0.1368	0.1757
Opacity	0.1334	0.1479	0.1751	0.1707	0.1604	0.1352	0.1049	0.1650	<u>0.1793</u>	0.1825
Atelectasis	0.6414	0.6578	0.7243	0.7034	0.6697	0.6483	0.4755	0.6993	0.7106	<u>0.7159</u>
Hyperinflation	0.3959	0.4022	0.4282	<u>0.4505</u>	0.4165	0.3645	0.1743	0.4304	0.4501	0.4628
Diffuse Fibrosis*	0.2105	0.2133	0.3875	0.3840	0.3293	0.1707	0.0050	<u>0.3917</u>	0.3602	0.4031
Surgical Clips	0.0469	0.0436	0.2229	0.1270	0.1819	0.0409	0.0269	0.0944	<u>0.2920</u>	0.3603
Low Lung Volumes	0.4624	0.4720	0.5111	<u>0.5258</u>	0.5170	0.4089	0.2274	0.5136	0.5208	0.5305
Aortic Aneurysm*	0.0077	0.0234	0.1010	0.1217	0.0778	0.0064	0.0043	0.1132	<u>0.1589</u>	0.1846
Pneumonia	0.3045	0.3679	0.4704	0.4476	0.3944	0.3430	0.2012	0.4396	0.4585	<u>0.4695</u>
Consolidation	0.3330	0.3765	0.4532	0.4340	0.3944	0.3471	0.2474	0.4229	0.4367	<u>0.4469</u>
Free Subdiaphragmatic Air*	0.0181	0.0293	0.1464	0.1476	0.0770	0.0118	0.0052	0.2848	<u>0.3195</u>	0.5151
Thoracic Spine Degenerative Changes*	0.0193	0.0199	0.0260	0.0274	0.0183	0.0165	0.0062	0.0276	0.0478	<u>0.0434</u>

Continued on next page

Abnormality	EVA-Base	RAD-DINO	ARK	CheXFound	BiomedCLIP	MedKLIP	BioViL-T	Med-CLIP	Med-AR-2B (ours)	Med-AR-8B (ours)
Port-a-cath [†]	0.1292	0.2633	0.4629	0.4219	0.4336	0.0299	0.0155	0.4716	<u>0.5031</u>	0.5291
Fibrosis	0.1511	0.2001	<u>0.3441</u>	0.3357	0.2472	0.1443	0.0918	0.3127	0.3278	0.3478
Lung Mass [†]	0.0726	0.1562	0.3144	0.2617	0.2027	0.0519	0.0195	0.2698	<u>0.3075</u>	0.3058
Endotracheal Tube	0.6125	0.6637	0.7116	0.7063	0.7063	0.6301	0.5439	0.6814	<u>0.7463</u>	0.7660
Pulmonary Arterial Hypertension Signs [†]	0.0273	0.0412	0.0757	<u>0.1364</u>	0.0614	0.0263	0.0174	0.1348	0.1187	0.1734
Effusion	0.8306	0.8452	0.8699	<u>0.8674</u>	0.8315	0.8326	0.6480	0.8589	0.8648	0.8668
Calcification	0.2096	0.2263	0.3268	0.3681	0.2888	0.1772	0.1139	0.3885	<u>0.4000</u>	0.4273
Degenerative Spine Changes [†]	0.0270	0.0281	0.0313	0.0362	0.0294	0.0179	0.0071	0.0313	<u>0.0366</u>	0.0383

S5.3 Sensitivity

Table S9: Per-label **Sensitivity** comparison across baseline, contrastive, and autoregressive models on the MIMIC dataset. Tail abnormalities are marked with an asterisk (*), and medium-prevalence abnormalities are marked with a dagger (†). For each row, the best value is shown in bold and the second-best value is underlined.

Abnormality	EVA-Base	RAD-DINO	ARK	CheXFound	BiomedCLIP	MedKLIP	BioViL-T	Med-CLIP	Med-AR-2B (ours)	Med-AR-8B (ours)
Diaphragm elevation	0.7113	0.7307	0.7942	0.8108	0.7555	0.7320	0.9171	0.8177	0.7887	<u>0.8522</u>
Parenchymal opacities†	0.8623	<u>0.8922</u>	0.8802	0.9222	<u>0.8922</u>	0.8144	0.7605	0.8443	0.8802	0.8323
Nodules	0.7195	0.6697	0.6868	<u>0.7274</u>	<u>0.7235</u>	0.6776	0.5531	0.6239	0.6645	0.7471
Pleural effusion	0.8516	0.8709	0.8747	0.8538	<u>0.8735</u>	0.8706	0.7845	0.8671	0.8535	0.8665
Congestive heart failure*	0.8095	0.8254	0.8571	0.8730	0.7619	0.9206	0.9206	<u>0.8889</u>	0.8413	0.8571
Support devices	0.9427	0.9301	0.9516	0.9498	0.9301	0.9409	0.9480	0.9570	<u>0.9642</u>	0.9749
PICC line	0.7813	0.8960	<u>0.9595</u>	0.9538	0.9143	0.7861	0.7813	0.8796	0.9566	0.9624
Pleural thickening	0.6724	<u>0.7966</u>	0.7483	0.7414	0.6776	0.6138	0.7845	0.7914	0.7879	0.8000
Soft tissue swelling†	0.6955	0.6925	0.7910	0.7522	0.8000	0.6537	0.8448	0.7612	<u>0.8209</u>	0.8149
Neoplastic nodules*	0.7604	0.7917	0.9271	0.8646	0.7292	0.7188	0.9792	<u>0.9375</u>	0.8750	0.8854
Hemothorax*	0.8140	0.8837	<u>0.9302</u>	0.8140	0.8721	0.8372	0.8721	0.8605	0.9419	0.8605
CABG†	0.7444	0.9000	0.9074	0.8889	0.8519	0.7148	0.7370	0.9111	<u>0.9333</u>	0.9370
Scarring*	<u>0.8841</u>	0.7246	0.8261	0.7536	0.7826	0.8406	0.7826	0.7971	0.9130	0.7681
Aortic tortuosity	0.7877	0.7777	0.7988	<u>0.8290</u>	0.7998	0.8390	0.5493	0.7988	<u>0.8290</u>	0.8390
Solitary nodules†	0.7518	<u>0.8540</u>	0.7591	0.7372	0.6788	0.9197	<u>0.8540</u>	0.7153	0.7664	0.6569
Tracheal deviation*	0.7927	0.8171	0.8049	0.6463	0.5610	0.9512	<u>0.9146</u>	0.8293	0.7561	0.7805
Scoliosis†	0.7204	<u>0.8059</u>	0.7303	0.7730	0.7007	0.7533	0.5855	0.7467	0.7632	0.8717
Pulmonary edema	0.8645	0.8352	<u>0.8822</u>	0.8778	0.8341	0.8290	0.8157	0.8724	0.8664	0.8854
Rib fractures	0.6524	0.6747	0.7945	0.8459	0.5908	0.6815	<u>0.8339</u>	0.7842	0.7380	0.7568
ARDS*	0.9176	0.9176	0.9176	0.8353	0.8824	<u>0.9294</u>	0.8471	0.9529	0.9059	0.8588
Aspiration†	0.7408	0.7824	<u>0.8264</u>	0.7359	0.7090	0.8166	0.8655	0.7408	0.7482	0.7800
Flattened diaphragm†	0.8917	0.9045	<u>0.9108</u>	0.8599	0.8471	0.8726	0.8471	0.9299	0.8217	0.8917
Reticular opacities†	0.7939	0.8702	0.8397	0.7786	0.7634	0.7405	0.8092	<u>0.8855</u>	0.9237	0.8321
Kyphosis*	0.7590	<u>0.8916</u>	0.9518	0.7590	0.8675	0.8193	0.7711	0.8313	0.8434	0.7711
Vascular redistribution cephalization	<u>0.8651</u>	0.7933	0.8083	0.8225	0.8554	0.7941	0.8394	0.8705	0.8571	0.8403
Degenerative changes†	<u>0.8526</u>	0.7649	0.8140	0.8105	0.8316	0.7544	0.7018	0.8211	<u>0.8526</u>	0.8667
Dialysis catheter*	0.7037	0.8519	0.9259	0.8765	0.8395	0.7654	0.9877	0.9136	0.9259	<u>0.9630</u>
Pigtail catheter*	0.8642	0.8765	0.9198	<u>0.9259</u>	0.8642	0.8086	0.7469	0.8889	0.8951	0.9321
Hilar lymphadenopathy†	0.7241	0.6207	0.8276	0.7438	0.7340	0.7438	0.7931	<u>0.8128</u>	0.7537	0.7685
Chest tube removal*	0.7671	<u>0.9041</u>	0.8219	<u>0.9041</u>	0.7397	0.9589	0.5205	0.8904	0.8904	0.7808
Retrocardiac opacity†	0.7374	0.7765	0.7039	0.6872	0.9050	0.6872	<u>0.7933</u>	0.7207	0.7263	0.7095
Pneumothorax	0.8103	0.8670	0.8876	0.8639	0.8557	0.8206	0.6649	0.8784	0.8753	<u>0.8845</u>
Vascular congestion†	0.7682	0.8545	<u>0.8636</u>	0.8000	0.7682	0.8136	0.7318	0.8409	0.9273	0.8455
Esophageal drainage tube*	0.9620	1.0000	0.9747	<u>0.9873</u>	0.9367	0.9747	0.7089	0.9620	0.9747	1.0000
Perihilar opacities	0.7681	0.7504	0.8319	<u>0.8053</u>	0.7965	0.7504	0.7327	0.7133	0.7416	0.7894
Interstitial lung disease pattern	0.7631	<u>0.7649</u>	0.7592	0.7584	0.6228	0.6854	0.7462	0.7336	0.7718	0.7475
Extubation*	0.9863	0.8356	0.8904	0.7123	0.8082	0.8082	0.7808	0.8630	<u>0.9452</u>	0.8356
Sternotomy wires	0.9647	0.9775	<u>0.9952</u>	<u>0.9952</u>	0.9759	0.9599	0.7095	0.9888	0.9936	0.9968
MPA enlargement†	0.8323	0.7081	0.7826	0.7702	0.6708	0.5901	<u>0.7950</u>	0.7081	0.6584	0.7081
Subcutaneous emphysema†	0.7826	0.8509	0.9379	0.9379	0.8696	0.7329	0.8323	<u>0.9130</u>	0.9006	0.9379
Bronchial wall thickening*	0.6857	0.6190	0.7238	0.6762	<u>0.7714</u>	0.7333	0.6286	0.7048	0.7905	0.7429
Chest wall deformity†	0.5939	0.7259	0.7360	0.8071	0.6599	0.7462	0.7919	0.8629	0.8274	<u>0.8376</u>
Feeding tube	0.9699	0.9826	0.9873	0.9953	0.9858	0.9731	0.9177	0.9905	0.9889	<u>0.9937</u>
Fluid overload*	0.7522	0.8584	0.9381	0.8230	0.8230	<u>0.8850</u>	0.8142	0.9381	0.8053	0.7434
Sternotomy†	0.9501	0.9948	<u>0.9895</u>	0.9948	0.9869	0.9423	0.8451	0.9869	0.9869	<u>0.9895</u>

Continued on next page

Abnormality	EVA-Base	RAD-DINO	ARK	CheXFound	BiomedCLIP	MedKLIP	BioViL-T	Med-CLIP	Med-AR-2B (ours)	Med-AR-8B (ours)
Granuloma [†]	0.8019	0.6415	0.6509	0.7783	0.6179	0.5849	0.7689	0.7264	0.7594	<u>0.7830</u>
Mediastinal mass*	<u>0.8450</u>	0.7132	0.7519	0.6512	0.6589	0.7519	0.8605	0.6899	0.6977	0.7829
Artifacts	0.7089	0.6509	0.7640	0.7139	<u>0.7404</u>	0.6942	0.6785	0.7325	0.6696	0.6775
Spinal hardware*	0.4098	0.5902	<u>0.8525</u>	0.6721	0.6557	0.3279	0.9180	0.6885	0.7377	0.7869
Cardiomegaly	0.8108	0.8111	0.8124	0.7796	0.7787	0.8048	<u>0.8306</u>	0.7998	0.8338	0.7971
Solitary pulmonary nodule [†]	0.7515	<u>0.8121</u>	0.7152	0.8000	0.7515	0.7697	0.7697	0.8909	0.7455	0.6545
Empyema*	0.8689	0.9344	0.8852	0.8525	0.7869	0.8689	0.8689	0.9672	<u>0.9508</u>	0.9016
Mediastinal widening	0.7495	0.6673	0.8150	<u>0.8224</u>	0.7701	0.7140	0.7682	0.8187	0.7607	0.8579
Chest tube	0.8490	0.8921	0.9461	<u>0.9384</u>	0.9076	0.8983	0.8382	0.9168	0.9461	0.9307
Malignant nodules*	0.8230	0.8230	0.9027	0.8230	0.7965	0.7699	0.7876	0.8230	0.8407	<u>0.8938</u>
Pacemaker	0.9592	0.9796	0.9833	<u>0.9870</u>	0.9852	0.8386	0.7978	<u>0.9870</u>	0.9907	0.9907
Cancer	0.7162	0.7582	0.8331	0.7832	0.7622	0.7595	0.8265	0.8108	0.8384	<u>0.8357</u>
Tracheostomy tube [†]	0.8688	0.8688	0.9462	<u>0.9634</u>	0.9140	0.8559	0.7892	0.9548	0.9570	0.9720
Atrial dilation*	0.7931	0.6897	0.7241	0.7241	0.6552	0.3793	0.9828	0.7069	0.7759	<u>0.8276</u>
MPA dilation [†]	0.6535	0.7008	<u>0.8110</u>	0.7874	0.7559	0.7559	0.5748	0.8346	0.7480	0.7795
Emphysema [†]	0.8962	0.7877	0.8750	0.8750	0.8585	0.8396	0.7123	<u>0.8821</u>	0.8349	0.8962
Interstitial edema	0.8768	0.8594	0.8272	0.8346	0.8456	0.9035	0.7914	<u>0.8851</u>	0.8483	0.8171
Diaphragmatic hernia [†]	0.7413	0.8297	0.8391	0.8328	0.7603	0.6719	0.8801	<u>0.8486</u>	0.8360	0.8454
Cavitary nodule*	0.7627	0.8305	<u>0.8983</u>	<u>0.8983</u>	0.7119	0.8305	0.6610	0.8136	0.8644	0.9153
Pericardial effusion [†]	0.6951	0.7195	0.8293	0.7683	0.8293	<u>0.8841</u>	0.7073	0.8598	0.8902	0.8293
Healed rib fracture [†]	0.6943	0.5886	0.8086	0.8343	0.7314	0.6771	0.7800	<u>0.8171</u>	0.7943	0.8086
Alveolar edema [†]	0.8354	0.9304	<u>0.9051</u>	0.8481	0.8418	0.8165	0.7785	0.8291	0.8291	0.8987
Aortic valve replacement*	0.9036	0.9398	<u>0.9639</u>	0.9398	0.8916	0.9398	0.8193	0.9036	0.9277	0.9759
Hilar enlargement	0.7042	0.7967	0.7151	<u>0.7949</u>	0.7840	0.6588	0.7659	0.7858	0.7151	0.7532
Hiatal hernia*	0.7273	0.8333	0.8788	0.8788	0.7879	0.8636	0.9697	0.8788	0.8939	<u>0.9242</u>
Bronchovascular crowding*	0.7714	0.8000	0.8381	0.7429	0.8190	0.7714	0.7429	0.8762	0.8095	<u>0.8571</u>
Internal jugular line [†]	0.8974	0.9475	<u>0.9761</u>	0.9666	0.9236	0.9260	0.7709	0.9594	0.9666	0.9809
Normal	0.8509	0.8737	0.8777	0.8742	0.8667	0.8440	0.8326	0.8600	<u>0.8754</u>	0.8475
Diffused nodules*	0.8362	0.7155	0.9052	<u>0.8707</u>	0.7586	0.7586	0.7672	0.7931	0.7845	0.8534
Calcified nodules*	0.9273	0.6727	0.7273	0.7636	0.5818	0.8000	<u>0.8364</u>	0.6727	0.7273	0.7091
Bibasilar opacities [†]	0.7832	0.8182	0.7902	0.7552	0.6853	0.7832	0.8322	0.8112	<u>0.8462</u>	0.8601
Clavicle fracture*	0.8115	0.5738	0.6148	0.5902	<u>0.7295</u>	0.8115	0.5492	0.6967	<u>0.7295</u>	0.5574
Nasogastric tube	0.9752	0.9752	0.9797	0.9955	0.9842	0.9744	0.9285	0.9902	0.9910	<u>0.9947</u>
Central venous line [†]	0.8633	0.9226	<u>0.9294</u>	0.9271	0.8884	0.7745	0.7585	0.9226	0.9362	0.8975
Left subclavian line*	<u>0.9367</u>	0.9114	<u>0.9367</u>	0.9494	0.7975	0.8987	0.7215	0.8734	0.8987	0.9494
Pneumomediastinum*	0.6400	0.6133	0.8933	0.8933	0.6667	0.6267	0.1867	0.7067	0.7333	<u>0.7733</u>
Interstitial thickening [†]	0.8088	0.7959	0.8140	0.8114	0.7726	0.7416	0.8450	<u>0.8398</u>	0.8243	0.7494
Dobhoff tube [†]	0.9327	0.9279	0.9279	0.9808	0.9663	0.8990	0.9567	<u>0.9760</u>	0.9567	0.9712
Bronchiectasis*	0.7662	<u>0.8831</u>	0.8312	0.8442	0.8182	0.7792	0.7532	0.7273	0.8312	0.9481
Swan-Ganz catheter [†]	0.9372	0.9324	<u>0.9565</u>	0.9420	0.9179	0.9275	0.8116	0.9517	0.9710	0.9710
Bullae*	0.5574	0.8361	0.8033	0.8033	0.8361	0.6721	1.0000	0.8525	0.8033	<u>0.8689</u>
Adenocarcinoma*	0.6731	0.8269	0.7885	0.7692	<u>0.8654</u>	0.7308	0.9808	0.8462	<u>0.8654</u>	0.7885
Mediastinal shift [†]	0.7989	0.8370	0.8587	0.9076	0.8478	0.8207	0.7554	<u>0.9022</u>	0.8859	0.8696
Mediastinal clips*	0.9174	0.9256	0.9917	0.9669	0.9587	0.9339	0.8347	<u>0.9752</u>	0.9504	<u>0.9752</u>
Opacity	0.7890	0.8434	<u>0.7979</u>	0.7673	0.7942	0.7785	0.7808	0.7658	0.7927	0.7450
Atelectasis	0.7351	0.7431	0.8046	<u>0.8012</u>	0.7599	0.7861	0.7644	0.7760	0.7976	0.7926
Hyperinflation	0.8410	0.8285	0.8420	<u>0.8524</u>	0.8337	0.8347	0.7983	0.8857	0.8462	0.8472
Diffuse fibrosis*	0.9508	0.9672	0.9672	0.9672	1.0000	0.9344	0.8197	0.9672	<u>0.9836</u>	0.9672
Surgical clips	0.4989	0.6518	0.7707	0.7219	0.7219	0.6539	0.7749	0.6985	<u>0.7877</u>	0.8323
Low lung volumes	0.7881	0.7602	0.7849	0.7985	0.8175	0.7255	0.8253	0.7838	0.8074	<u>0.8185</u>
Aortic aneurysm*	<u>0.8952</u>	0.5429	0.6381	0.8857	0.7238	0.3714	0.9524	0.7905	0.6000	0.6095
Pneumonia	0.7414	<u>0.7885</u>	0.7646	0.7708	0.7028	0.7680	0.7728	0.8046	0.7602	0.7667
Consolidation	0.7961	0.8431	0.8122	0.8421	0.7850	0.8075	0.7020	0.8116	0.7843	<u>0.8425</u>
Free subdiaphragmatic air*	0.6735	0.8469	0.8367	0.8469	0.7551	0.7653	0.9388	0.7347	0.8163	<u>0.8571</u>

Continued on next page

Abnormality	EVA-Base	RAD-DINO	ARK	CheXFound	BiomedCLIP	MedKLIP	BioViL-T	Med-CLIP	Med-AR-2B (ours)	Med-AR-8B (ours)
Thoracic spine degenerative changes*	<u>0.8785</u>	0.7850	0.7664	0.8318	0.8411	<u>0.8785</u>	0.7570	0.8224	<u>0.8785</u>	0.8879
Port-a-cath [†]	0.7187	0.8412	0.9638	0.9471	0.9415	0.7382	0.8162	0.9248	<u>0.9777</u>	0.9861
Fibrosis	0.7458	0.8246	<u>0.8473</u>	0.7745	0.7542	0.7733	0.6313	0.8067	0.8365	0.8580
Lung mass [†]	0.7228	0.7636	0.8804	0.8533	0.7745	0.6902	0.7772	0.8451	<u>0.8560</u>	0.8370
Endotracheal tube	0.9787	0.9813	0.9750	<u>0.9849</u>	0.9802	0.9704	0.9496	0.9782	0.9818	0.9854
Pulmonary arterial hypertension signs [†]	0.7600	0.7120	0.6560	0.7440	0.8400	0.7920	0.6640	0.6560	0.6920	<u>0.8240</u>
Effusion	0.8429	0.8601	0.8669	0.8494	0.8336	<u>0.8678</u>	0.7951	0.8414	0.8590	0.8744
Calcification	0.7545	0.7243	<u>0.8220</u>	0.8062	0.7624	0.7294	0.7897	0.7732	0.7947	0.8270
Degenerative spine changes [†]	0.7630	0.8519	0.8593	0.9333	0.8000	0.8593	0.3259	<u>0.9259</u>	0.9185	0.8963

S5.4 Specificity

Table S10: Per-label **Specificity** comparison across baseline, contrastive, and autoregressive models on the MIMIC dataset. Tail abnormalities are marked with an asterisk (*), and medium-prevalence abnormalities are marked with a dagger (†). For each row, the best value is shown in bold and the second-best value is underlined.

Abnormality	EVA-Base	RAD-DINO	ARK	CheXFound	BiomedCLIP	MedKLIP	BioViL-T	Med-CLIP	Med-AR-2B (ours)	Med-AR-8B (ours)
Diaphragm elevation	0.7651	0.7826	0.7739	0.8533	0.8421	0.5545	0.2131	0.8440	0.9018	0.8586
Parenchymal opacities†	0.7392	0.7465	<u>0.8114</u>	0.7207	0.7516	0.7588	0.6673	0.7963	0.7685	0.8239
Nodules	0.6128	0.7120	<u>0.8134</u>	0.7440	0.6781	0.6684	0.5918	0.8142	0.7951	0.7553
Pleural effusion	0.8140	0.8133	0.8389	0.8584	0.7923	0.8009	0.7095	0.8325	<u>0.8493</u>	0.8377
Congestive heart failure*	0.7424	0.7545	0.7655	0.6752	0.8219	0.5473	0.4684	0.7504	<u>0.7846</u>	0.7823
Support devices	0.7846	0.8139	0.8325	0.8189	<u>0.8283</u>	0.7751	0.7241	0.8077	0.8251	0.8110
PICC line	0.6081	0.7148	<u>0.9348</u>	0.9337	0.9065	0.5338	0.5221	0.8650	0.9320	0.9350
Pleural thickening	0.7081	0.6473	<u>0.7937</u>	0.8046	0.7616	0.7330	0.3895	0.7374	0.7692	0.7650
Soft tissue swelling†	0.8486	<u>0.9401</u>	0.9206	0.9586	0.8564	0.8323	0.3187	0.9388	0.9007	0.9167
Neoplastic nodules*	0.7633	0.8222	<u>0.9142</u>	0.9039	0.9316	0.7886	0.1045	0.7535	0.8911	0.9106
Hemothorax*	0.7609	0.7517	0.8081	<u>0.8578</u>	0.8450	0.5879	0.4819	0.8518	0.8375	0.8835
CABG†	0.8938	0.8450	0.8118	0.8687	0.8671	<u>0.8816</u>	0.4189	0.8329	0.8214	0.8368
Scarring*	0.5164	0.6971	0.7053	0.7956	0.6177	0.5741	0.2735	0.7297	0.6508	<u>0.7658</u>
Aortic tortuosity	0.7003	0.7497	0.7734	0.7805	0.7468	0.6102	0.6842	0.8065	0.7728	<u>0.7908</u>
Solitary nodules†	0.6878	0.6143	0.7625	0.7465	0.7611	0.5525	0.5337	<u>0.7808</u>	0.7484	0.8186
Tracheal deviation*	0.5903	0.5326	0.5942	<u>0.6777</u>	0.7270	0.3436	0.2655	0.5713	0.6279	0.6431
Scoliosis†	0.7702	0.6604	0.7788	0.9085	0.8497	0.6189	0.6269	<u>0.8794</u>	0.8406	0.8277
Pulmonary edema	0.7530	<u>0.7881</u>	0.7769	0.7738	0.7864	0.7905	0.6752	0.7741	0.7841	0.7686
Rib fractures	0.6340	0.6834	0.8215	0.7343	0.7847	0.5720	0.2848	0.8017	<u>0.8476</u>	0.8738
ARDS*	0.8507	0.8315	0.8911	<u>0.9165</u>	0.8683	0.8764	0.7415	0.8159	0.8627	0.9179
Aspiration†	0.5870	0.5767	0.5968	0.7129	0.6659	0.4923	0.3466	<u>0.6890</u>	0.6814	0.6658
Flattened diaphragm†	0.7950	0.8115	0.8115	0.8676	<u>0.8697</u>	0.7969	0.5344	<u>0.8046</u>	0.8899	0.8572
Reticular opacities†	0.7315	0.7146	0.8336	0.8590	0.8211	0.7922	0.3387	0.7737	0.7364	<u>0.8427</u>
Kyphosis*	0.8499	0.7766	0.7186	<u>0.8877</u>	0.7846	0.7650	0.3191	0.8704	0.8751	0.9035
Vascular redistribution cephalization	0.5603	<u>0.6453</u>	0.6679	0.6445	0.5916	0.6403	0.4654	0.5857	0.6025	0.6325
Degenerative changes†	0.6020	<u>0.6811</u>	0.6579	0.6870	0.6619	0.6363	0.4163	0.6523	0.6477	0.6375
Dialysis catheter*	0.7521	0.8832	0.8959	<u>0.9418</u>	0.9258	0.4998	0.0325	0.9079	0.9250	0.9455
Pigtail catheter*	0.6281	0.6680	0.9515	0.8651	0.8667	0.6300	0.5015	0.8854	<u>0.9533</u>	0.9574
Hilar lymphadenopathy†	0.6161	<u>0.8011</u>	0.7192	0.7857	0.7579	0.5807	0.2793	0.6891	0.7640	0.8355
Chest tube removal*	0.7331	0.7029	0.8521	0.7975	<u>0.8661</u>	0.3142	0.6211	0.7715	0.8362	0.9171
Retrocardiac opacity†	0.5999	0.5495	0.7234	<u>0.7377</u>	0.4655	0.5877	0.4698	0.6818	0.6462	0.7484
Pneumothorax	0.7576	0.7783	0.9288	0.9003	0.8371	0.7297	0.6483	0.8586	0.8880	<u>0.9096</u>
Vascular congestion†	0.6110	0.5635	0.6165	0.6334	<u>0.6305</u>	0.6044	0.6120	0.5939	0.5117	0.6223
Esophageal drainage tube*	0.7690	0.7940	<u>0.8145</u>	0.7987	0.7952	0.7890	0.6465	0.8078	0.8214	0.8128
Perihilar opacities	0.6700	0.7118	0.6964	0.7079	0.6918	0.6942	0.5642	0.7816	<u>0.7722</u>	0.7511
Interstitial lung disease pattern	0.6311	0.6629	0.7429	0.7351	0.8077	0.7182	0.4358	0.7566	0.7196	<u>0.7621</u>
Extubation*	0.4861	0.7078	0.6261	<u>0.7550</u>	0.6998	0.7180	0.5285	0.7209	0.7375	0.8153
Sternotomy wires	0.8919	<u>0.9095</u>	0.9019	0.9058	0.9106	0.8822	0.5645	0.9084	0.9062	0.9062
MPA enlargement†	0.4746	0.6539	0.6304	0.7763	0.7478	0.7267	0.4046	0.7984	<u>0.8406</u>	0.8868
Subcutaneous emphysema†	0.8549	0.9194	<u>0.9603</u>	0.9476	0.9245	0.7863	0.4162	0.9388	0.9582	0.9623
Bronchial wall thickening*	0.6969	<u>0.7611</u>	0.7577	0.7775	0.5335	0.5394	0.5648	0.7448	0.6565	0.6937
Chest wall deformity†	<u>0.8057</u>	0.7055	0.8152	0.7587	0.7536	0.5866	0.4014	0.7182	0.7694	0.7966
Feeding tube	0.7983	0.8040	0.8182	0.8170	0.8162	0.8112	0.7971	0.8146	0.8284	<u>0.8204</u>
Fluid overload*	<u>0.7029</u>	0.5939	0.5637	0.6128	0.6454	0.5800	0.5629	0.5557	0.6929	0.7470
Sternotomy†	0.8791	0.8909	0.8916	0.8937	0.8902	0.8746	0.3595	<u>0.8948</u>	0.8958	0.8958

Continued on next page

Abnormality	EVA-Base	RAD-DINO	ARK	CheXFound	BiomedCLIP	MedKLIP	BioViL-T	Med-CLIP	Med-AR-2B (ours)	Med-AR-8B (ours)
Granuloma [†]	0.5792	0.6937	0.8867	0.6844	0.6965	0.6996	0.4879	0.7799	0.8049	<u>0.8232</u>
Mediastinal mass*	0.4966	0.6780	0.7992	0.8578	0.7818	0.4914	0.2051	0.8318	0.8530	0.8062
Artifacts	0.5887	0.6316	0.5790	0.6332	0.5994	0.5350	0.5055	0.6048	0.6973	<u>0.6869</u>
Spinal hardware*	0.8491	0.6141	0.9187	0.9226	0.7689	0.8303	0.2548	0.8173	<u>0.9547</u>	0.9641
Cardiomegaly	0.7150	0.6962	0.7595	0.7887	0.7628	0.6762	0.4889	0.7613	0.7380	<u>0.7786</u>
Solitary pulmonary nodule [†]	0.7156	0.6948	<u>0.7804</u>	0.6774	0.6524	0.6851	0.5237	0.5490	0.7196	0.8267
Emphyema*	0.7212	0.7564	0.8239	0.9065	0.8440	0.6167	0.5265	0.8394	0.8472	<u>0.8950</u>
Mediastinal widening	0.7212	<u>0.7840</u>	0.7677	0.7564	0.7342	0.6845	0.5364	0.7666	0.8274	0.7591
Chest tube	0.7712	0.8578	<u>0.9320</u>	0.8917	0.8761	0.6564	0.5401	0.8804	0.9062	0.9399
Malignant nodules*	0.6292	0.7887	0.8920	<u>0.9020</u>	0.8267	0.7172	0.4033	0.8553	0.9029	0.8548
Pacemaker	0.8986	0.9497	0.9437	0.9515	0.9420	0.8386	0.5273	0.9477	0.9492	<u>0.9500</u>
Cancer	0.7705	0.7807	0.7794	0.8346	0.8083	0.6525	0.3522	0.7963	0.7995	<u>0.8088</u>
Tracheostomy tube [†]	0.7711	0.8982	0.9550	<u>0.9792</u>	0.9406	0.6456	0.6518	0.9582	0.9860	0.9786
Atrial dilation*	0.6177	0.6702	<u>0.7099</u>	0.6905	0.7088	0.7534	0.0224	0.6529	0.6804	0.6208
MPA dilation [†]	0.7098	0.6975	0.6513	0.7710	0.6930	0.5819	0.5708	0.7576	<u>0.8247</u>	0.8697
Emphysema [†]	0.7087	<u>0.8441</u>	0.8135	0.8324	0.8102	0.7780	0.6731	0.8232	0.8606	0.8242
Interstitial edema	0.6327	0.6636	0.7452	0.7297	0.6804	0.6191	0.5698	0.6742	0.7107	<u>0.7305</u>
Diaphragmatic hernia [†]	0.7305	0.6515	0.7829	<u>0.8803</u>	0.7861	0.7100	0.2305	0.7264	0.8294	0.8896
Cavitary nodule*	0.7799	0.8737	0.7724	0.7957	<u>0.8635</u>	0.6949	0.6579	0.8737	0.7841	0.8025
Pericardial effusion [†]	0.7924	0.7681	0.7873	0.8483	0.7598	0.5552	0.5285	0.7799	0.7351	<u>0.8224</u>
Healed rib fracture [†]	0.6361	0.7888	<u>0.8664</u>	0.8094	0.6717	0.6569	0.3744	0.8383	0.8065	0.8939
Alveolar edema [†]	0.7846	0.7183	0.7762	0.8162	0.8158	<u>0.8332</u>	0.7166	0.8584	0.8166	0.7673
Aortic valve replacement*	0.8836	0.8573	0.8674	0.8636	0.8622	0.8689	0.2774	0.9028	0.8747	<u>0.8890</u>
Hilar enlargement	0.6341	0.6285	0.7790	0.7267	0.6832	0.7096	0.3793	0.7372	0.8012	<u>0.7888</u>
Hiatal hernia*	0.7909	0.8099	0.8088	<u>0.9255</u>	0.8486	0.6836	0.3774	0.8017	0.9030	0.9486
Bronchovascular crowding*	0.7926	0.8222	0.8160	0.8719	0.7811	0.7340	0.4688	0.7864	<u>0.8331</u>	0.7784
Internal jugular line [†]	0.7263	0.7283	0.8309	0.8650	0.9004	0.6754	0.7166	0.8483	<u>0.8825</u>	0.8808
Normal	0.8467	0.8409	0.8563	0.8648	0.8340	0.8547	0.7952	<u>0.8712</u>	0.8599	0.8830
Diffused nodules*	0.6552	0.8729	0.8631	0.8333	0.8756	0.8253	0.4406	0.8906	0.8981	<u>0.8928</u>
Calcified nodules*	0.2467	0.6020	0.8367	0.7368	0.7654	0.3156	0.2849	<u>0.8671</u>	0.8457	0.8924
Bibasilar opacities [†]	0.6381	0.6300	0.7090	<u>0.7161</u>	0.7438	0.5763	0.4184	0.6904	0.6267	0.6511
Clavicle fracture*	0.4042	0.6512	0.7543	0.8076	0.6478	0.4451	0.5716	<u>0.8108</u>	0.6398	0.8392
Nasogastric tube	0.8339	0.8486	0.8567	0.8491	0.8411	0.8421	0.8293	0.8462	<u>0.8501</u>	0.8490
Central venous line [†]	0.5697	0.7235	0.7776	0.7972	0.7941	0.5932	0.5618	0.7818	<u>0.8142</u>	0.8476
Left subclavian line*	0.6483	0.7623	0.8198	0.7979	0.9106	0.7293	0.4635	0.8040	<u>0.9151</u>	0.9163
Pneumomediastinum*	0.8616	0.8535	0.7878	0.7801	0.9146	0.7938	0.8492	0.8847	<u>0.8996</u>	0.8779
Interstitial thickening [†]	0.7352	0.7616	<u>0.8182</u>	0.7997	0.7770	0.7808	0.4283	0.7658	0.7841	0.8489
Dobhoff tube [†]	0.8220	0.8903	0.9465	0.8771	0.8787	0.8840	0.7629	0.8469	0.9279	<u>0.9456</u>
Bronchiectasis*	<u>0.8791</u>	0.7786	0.8703	0.8418	0.7981	0.8311	0.5779	0.9152	0.8653	0.7630
Swan-Ganz catheter [†]	0.8044	0.8897	0.9506	0.9309	<u>0.9621</u>	0.7473	0.4937	0.9063	0.9580	0.9901
Bullae*	<u>0.8401</u>	0.7448	0.7778	0.8233	0.7150	0.7628	0.0094	0.7720	0.7893	0.8543
Adenocarcinoma*	0.6782	0.6383	<u>0.7936</u>	0.7488	0.7157	0.5888	0.0527	0.7796	0.7262	0.8163
Mediastinal shift [†]	0.8228	0.8037	0.8378	0.8120	0.8335	0.6810	0.4550	0.8192	<u>0.8465</u>	0.9025
Mediastinal clips*	<u>0.9291</u>	0.8959	0.8797	0.9015	0.9070	0.8848	0.2924	0.8936	0.9297	0.8912
Opacity	0.5553	0.5286	0.6269	<u>0.6359</u>	0.5824	0.5851	0.4921	0.6212	0.5935	0.6649
Atelectasis	0.7054	0.7212	0.7100	0.7011	0.7065	0.6508	0.5465	0.7160	0.7056	<u>0.7174</u>
Hyperinflation	0.8058	0.8290	0.8232	0.8300	0.8345	0.8050	0.6565	0.7952	<u>0.8356</u>	0.8423
Diffuse fibrosis*	0.8818	0.9116	0.9415	0.9332	0.9180	0.8988	0.5378	0.9718	0.9443	<u>0.9686</u>
Surgical clips	0.7848	0.6267	0.7728	0.7229	0.7721	0.6119	0.4224	0.7331	<u>0.8260</u>	0.8720
Low lung volumes	0.7491	<u>0.7801</u>	0.7707	0.7693	0.7433	0.7613	0.4979	0.7825	0.7612	0.7620
Aortic aneurysm*	0.2585	0.8244	0.8438	0.6599	0.6752	0.7435	0.1242	0.6889	<u>0.8942</u>	0.9367
Pneumonia	0.6317	0.6396	0.7399	0.7157	<u>0.7392</u>	0.6325	0.4575	0.6633	0.7198	0.7282
Consolidation	0.6345	0.6394	<u>0.7232</u>	0.6761	0.6886	0.6401	0.6120	0.6955	0.7255	0.6783
Free subdiaphragmatic air*	0.7240	0.6598	0.7074	0.7410	0.8481	0.6010	0.1877	<u>0.8667</u>	0.8422	0.9123

Continued on next page

Abnormality	EVA-Base	RAD-DINO	ARK	CheXFound	BiomedCLIP	MedKLIP	BioViL-T	Med-CLIP	Med-AR-2B (ours)	Med-AR-8B (ours)
Thoracic spine degenerative changes*	0.6473	0.6871	0.7707	0.6768	0.6098	0.5664	0.4928	<u>0.6985</u>	0.6467	0.6778
Port-a-cath [†]	0.7700	0.8969	0.9303	0.9282	0.9349	0.5169	0.2368	0.9518	<u>0.9560</u>	0.9664
Fibrosis	0.7149	0.6950	0.7744	0.8397	0.7682	0.6766	0.6705	<u>0.7982</u>	0.7778	0.7715
Lung mass [†]	0.7304	<u>0.8325</u>	0.7803	0.8324	0.8314	0.7063	0.3664	0.8075	0.8062	0.8480
Endotracheal tube	0.8955	0.9068	0.9231	0.9262	0.9170	0.8978	0.8704	0.9226	<u>0.9375</u>	0.9387
Pulmonary arterial hypertension signs [†]	0.5727	0.7024	0.7740	0.7729	0.5885	0.5924	0.5498	0.8519	<u>0.8143</u>	0.7156
Effusion	0.8233	0.8244	0.8444	0.8618	0.8333	0.8035	0.6984	<u>0.8560</u>	0.8423	0.8305
Calcification	0.6772	0.7240	0.7043	0.7675	0.7165	0.6506	0.4755	0.8084	<u>0.7945</u>	0.7883
Degenerative spine changes [†]	0.6893	0.6877	0.7054	0.6631	0.7214	0.5777	<u>0.7146</u>	0.6375	0.6735	0.6978

S5.5 Youden’s J

Table S11: Per-label **Youden’s J Score** comparison across baseline, contrastive, and autoregressive models on the **MIMIC** dataset. Tail abnormalities are marked with an asterisk (*), and medium-prevalence abnormalities are marked with a dagger (†). For each row, the best value is shown in bold and the second-best value is underlined.

Abnormality	EVA-Base	RAD-DINO	ARK	CheXFound	BiomedCLIP	MedKLIP	BioViL-T	Med-CLIP	Med-AR-2B (ours)	Med-AR-8B (ours)
Diaphragm elevation	0.4764	0.5133	0.5681	0.6641	0.5976	0.2865	0.1302	0.6617	<u>0.6905</u>	0.7108
Parenchymal opacities†	0.6015	0.6387	0.6916	0.6429	0.6438	0.5732	0.4278	0.6406	0.6487	<u>0.6562</u>
Nodules	0.3323	0.3817	<u>0.5002</u>	0.4714	0.4016	0.3460	0.1449	0.4381	0.4596	0.5024
Pleural effusion	0.6656	0.6842	0.7136	<u>0.7122</u>	0.6658	0.6715	0.4940	0.6996	0.7028	0.7042
Congestive heart failure*	0.5519	0.5799	0.6226	0.5482	0.5838	0.4679	0.3890	<u>0.6393</u>	0.6259	0.6394
Support devices	0.7273	0.7440	0.7841	0.7687	0.7584	0.7160	0.6721	0.7647	0.7893	<u>0.7859</u>
PICC line	0.3894	0.6108	<u>0.8943</u>	0.8875	0.8208	0.3199	0.3034	0.7446	0.8886	0.8974
Pleural thickening	0.3805	0.4439	0.5420	0.5460	0.4392	0.3468	0.1740	0.5288	<u>0.5571</u>	0.5650
Soft tissue swelling†	0.5441	0.6326	0.7116	0.7108	0.6564	0.4860	0.1635	0.7000	<u>0.7216</u>	0.7316
Neoplastic nodules*	0.5237	0.6139	0.8413	0.7685	0.6608	0.5074	0.0837	0.6910	0.7661	<u>0.7960</u>
Hemothorax*	0.5749	0.6354	0.7383	0.6718	0.7171	0.4251	0.3540	0.7123	0.7794	<u>0.7440</u>
CABG†	0.6382	0.7450	0.7192	<u>0.7576</u>	0.7190	0.5964	0.1559	0.7440	0.7547	0.7738
Scarring*	0.4005	0.4217	0.5314	<u>0.5492</u>	0.4003	0.4147	0.0561	0.5268	0.5638	0.5339
Aortic tortuosity	0.4880	0.5274	0.5722	<u>0.6095</u>	0.5466	0.4492	0.2335	0.6053	0.6018	0.6298
Solitary nodules†	0.4396	0.4683	0.5216	0.4837	0.4399	0.4722	0.3877	0.4961	<u>0.5148</u>	0.4755
Tracheal deviation*	0.3830	0.3497	0.3991	0.3240	0.2880	0.2948	0.1801	<u>0.4006</u>	0.3840	0.4236
Scoliosis†	0.4906	0.4663	0.5091	<u>0.6815</u>	0.5504	0.3722	0.2124	0.6261	0.6038	0.6994
Pulmonary edema	0.6175	0.6233	0.6591	0.6516	0.6205	0.6195	0.4909	0.6465	0.6505	<u>0.6540</u>
Rib fractures	0.2864	0.3581	<u>0.6160</u>	0.5802	0.3755	0.2535	0.1187	0.5859	0.5856	0.6306
ARDS*	0.7683	0.7491	0.8087	0.7518	0.7507	<u>0.8058</u>	0.5886	0.7688	0.7686	0.7767
Aspiration†	0.3278	0.3591	0.4232	0.4488	0.3749	0.3089	0.2121	0.4298	0.4296	<u>0.4458</u>
Flattened diaphragm†	0.6867	0.7160	0.7223	0.7275	0.7168	0.6695	0.3815	<u>0.7345</u>	0.7116	0.7489
Reticular opacities†	0.5254	0.5848	<u>0.6733</u>	0.6376	0.5845	0.5327	0.1479	0.6592	0.6601	0.6748
Kyphosis*	0.6089	0.6682	0.6704	0.6467	0.6521	0.5843	0.0902	<u>0.7017</u>	0.7185	0.6746
Vascular redistribution cephalization	0.4254	0.4386	0.4762	0.4670	0.4470	0.4344	0.3048	0.4562	0.4596	<u>0.4728</u>
Degenerative changes†	0.4546	0.4460	0.4719	0.4975	0.4935	0.3907	0.1181	0.4734	<u>0.5003</u>	0.5042
Dialysis catheter*	0.4558	0.7351	0.8218	0.8183	0.7653	0.2652	0.0202	0.8215	<u>0.8509</u>	0.9085
Pigtail catheter*	0.4923	0.5445	<u>0.8713</u>	0.7910	0.7309	0.4386	0.2484	0.7743	0.8484	0.8895
Hilar lymphadenopathy†	0.3402	0.4218	<u>0.5468</u>	0.5295	0.4919	0.3245	0.0724	0.5019	0.5177	0.6040
Chest tube removal*	0.5002	0.6070	0.6740	<u>0.7016</u>	0.6058	0.2731	0.1416	0.6619	0.7266	0.6979
Retrocardiac opacity†	0.3373	0.3260	<u>0.4273</u>	0.4249	0.3705	0.2749	0.2631	0.4025	0.3725	0.4579
Pneumothorax	0.5679	0.6453	0.8164	0.7642	0.6928	0.5503	0.3132	0.7370	0.7633	<u>0.7941</u>
Vascular congestion†	0.3792	0.4180	0.4801	0.4334	0.3987	0.4180	0.3438	0.4348	0.4390	<u>0.4678</u>
Esophageal drainage tube*	0.7310	0.7940	0.7892	0.7860	0.7319	0.7637	0.3554	0.7698	<u>0.7961</u>	0.8128
Perihilar opacities	0.4381	0.4622	<u>0.5283</u>	0.5132	0.4883	0.4446	0.2969	0.4949	0.5138	0.5405
Interstitial lung disease pattern	0.3942	0.4278	<u>0.5021</u>	0.4935	0.4305	0.4036	0.1820	0.4902	0.4914	0.5096
Extubation*	0.4724	0.5434	0.5165	0.4673	0.5080	0.5262	0.3093	0.5839	0.6827	<u>0.6509</u>
Sternotomy wires	0.8566	0.8870	0.8971	<u>0.9010</u>	0.8865	0.8421	0.2740	0.8972	0.8998	0.9030
MPA enlargement†	0.3069	0.3620	0.4130	<u>0.5465</u>	0.4186	0.3168	0.1996	0.5065	0.4990	0.5949
Subcutaneous emphysema†	0.6375	0.7703	<u>0.8982</u>	0.8855	0.7941	0.5192	0.2485	0.8518	0.8588	0.9002
Bronchial wall thickening*	0.3826	0.3801	0.4815	<u>0.4537</u>	0.3049	0.2727	0.1934	0.4496	0.4470	0.4366
Chest wall deformity†	0.3996	0.4314	0.5512	0.5658	0.4135	0.3328	0.1933	0.5811	<u>0.5968</u>	0.6342
Feeding tube	0.7682	0.7866	0.8055	0.8123	0.8020	0.7843	0.7148	0.8051	0.8173	<u>0.8141</u>
Fluid overload*	0.4551	0.4523	0.5018	0.4358	0.4684	0.4650	0.3771	0.4938	<u>0.4982</u>	0.4904
Sternotomy†	0.8292	<u>0.8857</u>	0.8811	0.8885	0.8771	0.8169	0.2046	0.8817	0.8827	0.8853
Granuloma†	0.3811	0.3352	0.5376	0.4627	0.3144	0.2845	0.2568	0.5063	<u>0.5643</u>	0.6062

Continued on next page

Abnormality	EVA-Base	RAD-DINO	ARK	CheXFound	BiomedCLIP	MedKLIP	BioViL-T	Med-CLIP	Med-AR-2B (ours)	Med-AR-8B (ours)
Mediastinal mass*	0.3416	0.3912	<u>0.5511</u>	0.5090	0.4407	0.2433	0.0656	0.5217	0.5507	0.5891
Artifacts	0.2976	0.2825	0.3430	0.3471	0.3398	0.2292	0.1840	0.3373	0.3669	<u>0.3644</u>
Spinal hardware*	0.2589	0.2043	0.7712	0.5947	0.4246	0.1582	0.1728	0.5058	0.6924	<u>0.7510</u>
Cardiomegaly	0.5258	0.5073	<u>0.5719</u>	0.5683	0.5415	0.4810	0.3195	0.5611	0.5718	0.5757
Solitary pulmonary nodule [†]	0.4671	0.5069	<u>0.4956</u>	0.4774	0.4039	0.4548	0.2934	0.4399	0.4651	0.4812
Empyema*	0.5901	0.6908	0.7091	0.7590	0.6309	0.4856	0.3954	0.8066	<u>0.7980</u>	0.7966
Mediastinal widening	0.4707	0.4513	0.5827	0.5788	0.5043	0.3985	0.3046	0.5853	<u>0.5881</u>	0.6170
Chest tube	0.6202	0.7499	0.8781	0.8301	0.7837	0.5547	0.3783	0.7972	0.8523	<u>0.8706</u>
Malignant nodules*	0.4522	0.6117	0.7947	0.7250	0.6232	0.4871	0.1909	0.6783	0.7436	<u>0.7486</u>
Pacemaker	0.8578	0.9293	0.9270	0.9385	0.9272	0.6772	0.3251	0.9347	<u>0.9399</u>	0.9407
Cancer	0.4867	0.5389	0.6125	0.6178	0.5705	0.4120	0.1787	0.6071	<u>0.6379</u>	0.6445
Tracheostomy tube [†]	0.6399	0.7670	0.9012	0.9426	0.8546	0.5015	0.4410	0.9130	<u>0.9430</u>	0.9506
Atrial dilation*	0.4108	0.3599	0.4340	0.4146	0.3640	0.1327	0.0052	0.3598	0.4563	<u>0.4484</u>
MPA dilation [†]	0.3633	0.3983	0.4623	0.5584	0.4489	0.3378	0.1456	<u>0.5922</u>	0.5727	0.6492
Emphysema [†]	0.6049	0.6318	0.6885	<u>0.7074</u>	0.6687	0.6176	0.3854	0.7053	0.6955	0.7204
Interstitial edema	0.5095	0.5230	0.5724	<u>0.5643</u>	0.5260	0.5226	0.3612	0.5593	0.5590	0.5476
Diaphragmatic hernia [†]	0.4718	0.4812	0.6220	<u>0.7131</u>	0.5464	0.3819	0.1106	0.5750	0.6654	0.7350
Cavitary nodule*	0.5426	<u>0.7042</u>	0.6707	0.6940	0.5754	0.5254	0.3189	0.7015	0.6485	0.7178
Pericardial effusion [†]	0.4875	0.4876	0.6166	0.6166	0.5891	0.4393	0.2358	<u>0.6397</u>	0.6253	0.6517
Healed rib fracture [†]	0.3304	0.3774	<u>0.6750</u>	0.6437	0.4031	0.3340	0.1544	0.6554	0.6008	0.7025
Alveolar edema [†]	0.6200	0.6487	<u>0.6813</u>	0.6643	0.6576	0.6497	0.4951	0.6875	0.6457	0.6660
Aortic valve replacement*	0.7872	0.7971	<u>0.8313</u>	0.8034	0.7538	0.8087	0.0967	0.8064	0.8024	0.8649
Hilar enlargement	0.3383	0.4252	0.4941	0.5216	0.4672	0.3684	0.1452	<u>0.5230</u>	0.5163	0.5420
Hiatal hernia*	0.5182	0.6432	0.6876	<u>0.8043</u>	0.6365	0.5472	0.3471	0.6805	0.7969	0.8728
Bronchovascular crowding*	0.5640	0.6222	<u>0.6541</u>	0.6148	0.6001	0.5054	0.2117	0.6626	0.6426	0.6355
Internal jugular line [†]	0.6237	0.6758	0.8070	0.8316	0.8240	0.6014	0.4875	0.8077	<u>0.8491</u>	0.8617
Normal	0.6976	0.7146	0.7340	0.7390	0.7007	0.6987	0.6278	0.7312	<u>0.7353</u>	0.7305
Diffused nodules*	0.4914	0.5884	0.7683	0.7040	0.6342	0.5839	0.2078	0.6837	0.6826	<u>0.7462</u>
Calcified nodules*	0.1740	0.2747	0.5640	0.5004	0.3472	0.1156	0.1213	0.5398	<u>0.5730</u>	0.6015
Bibasilar opacities [†]	0.4213	0.4482	0.4992	0.4713	0.4291	0.3595	0.2506	<u>0.5016</u>	0.4729	0.5112
Clavicle fracture*	0.2157	0.2250	0.3691	<u>0.3978</u>	0.3773	0.2566	0.1208	0.5075	0.3693	0.3966
Nasogastric tube	0.8091	0.8238	0.8364	0.8446	0.8253	0.8165	0.7578	0.8364	0.8411	<u>0.8437</u>
Central venous line [†]	0.4330	0.6461	0.7070	0.7243	0.6825	0.3677	0.3203	0.7044	0.7504	<u>0.7451</u>
Left subclavian line*	0.5850	0.6737	0.7565	0.7473	0.7081	0.6280	0.1850	0.6774	<u>0.8138</u>	0.8657
Pneumomediastinum*	0.5016	0.4668	0.6811	<u>0.6734</u>	0.5813	0.4205	0.0359	0.5914	0.6329	0.6512
Interstitial thickening [†]	0.5440	0.5575	0.6322	<u>0.6111</u>	0.5496	0.5224	0.2733	0.6056	0.6084	0.5983
Dobhoff tube [†]	0.7547	0.8182	0.8744	0.8579	0.8450	0.7830	0.7196	0.8229	<u>0.8846</u>	0.9168
Bronchiectasis*	0.6453	0.6617	<u>0.7015</u>	0.6860	0.6163	0.6103	0.3311	0.6425	0.6965	0.7111
Swan-Ganz catheter [†]	0.7416	0.8221	0.9071	0.8729	0.8800	0.6748	0.3053	0.8580	<u>0.9290</u>	0.9611
Bullae*	0.3975	0.5809	0.5811	<u>0.6266</u>	0.5511	0.4349	0.0094	0.6245	0.5926	0.7232
Adenocarcinoma*	0.3513	0.4652	0.5821	0.5180	0.5811	0.3196	0.0335	0.6258	0.5916	<u>0.6048</u>
Mediastinal shift [†]	0.6217	0.6407	0.6965	0.7196	0.6813	0.5017	0.2104	0.7214	<u>0.7324</u>	0.7721
Mediastinal clips*	0.8465	0.8215	<u>0.8714</u>	0.8684	0.8657	0.8187	0.1271	0.8688	0.8801	0.8664
Opacity	0.3443	0.3720	0.4248	0.4032	0.3766	0.3636	0.2729	0.3870	0.3862	<u>0.4099</u>
Atelectasis	0.4405	0.4643	0.5146	0.5023	0.4664	0.4369	0.3109	0.4920	0.5032	<u>0.5100</u>
Hyperinflation	0.6468	0.6575	0.6652	<u>0.6824</u>	0.6682	0.6397	0.4548	0.6809	0.6818	0.6895
Diffuse fibrosis*	0.8326	0.8788	0.9087	0.9004	0.9180	0.8332	0.3575	0.9390	0.9279	<u>0.9358</u>
Surgical clips	0.2837	0.2785	0.5435	0.4448	0.4940	0.2658	0.1973	0.4316	<u>0.6137</u>	0.7043
Low lung volumes	0.5372	0.5403	0.5556	0.5678	0.5608	0.4868	0.3232	0.5663	<u>0.5686</u>	0.5805
Aortic aneurysm*	0.1537	0.3673	0.4819	<u>0.5456</u>	0.3990	0.1149	0.0766	0.4794	0.4942	0.5462
Pneumonia	0.3731	0.4281	0.5045	0.4865	0.4420	0.4005	0.2303	0.4679	0.4800	<u>0.4949</u>
Consolidation	0.4306	0.4825	0.5354	0.5182	0.4736	0.4476	0.3140	0.5071	0.5098	<u>0.5208</u>
Free subdiaphragmatic air*	0.3975	0.5067	0.5441	0.5879	0.6032	0.3663	0.1265	0.6014	<u>0.6585</u>	0.7694
Thoracic spine degenerative changes*	0.5258	0.4721	<u>0.5371</u>	0.5086	0.4509	0.4449	0.2498	0.5209	0.5252	0.5657

Continued on next page

Abnormality	EVA-Base	RAD-DINO	ARK	CheXFound	BiomedCLIP	MedKLIP	BioViL-T	Med-CLIP	Med-AR-2B (ours)	Med-AR-8B (ours)
Port-a-cath [†]	0.4887	0.7381	0.8941	0.8753	0.8764	0.2551	0.0530	0.8766	<u>0.9337</u>	0.9525
Fibrosis	0.4607	0.5196	<u>0.6217</u>	0.6142	0.5224	0.4499	0.3018	0.6049	0.6143	0.6295
Lung mass [†]	0.4532	0.5961	0.6607	0.6857	0.6059	0.3965	0.1436	0.6526	0.6622	<u>0.6850</u>
Endotracheal tube	0.8742	0.8881	0.8981	0.9111	0.8972	0.8682	0.8200	0.9008	<u>0.9193</u>	0.9241
Pulmonary arterial hypertension signs [†]	0.3327	0.4144	0.4300	<u>0.5169</u>	0.4285	0.3844	0.2138	0.5079	0.5063	0.5396
Effusion	0.6662	0.6845	0.7113	<u>0.7112</u>	0.6669	0.6713	0.4935	0.6974	0.7013	0.7049
Calcification	0.4317	0.4483	0.5263	0.5737	0.4789	0.3800	0.2652	0.5816	<u>0.5892</u>	0.6153
Degenerative spine changes [†]	0.4523	0.5396	0.5647	0.5964	0.5214	0.4370	0.0405	0.5634	0.5920	<u>0.5941</u>

S6 PadChest: Per-Label Results

The following tables report the existing results for 29 labels, in the order AUROC, AUPRC, sensitivity, specificity, and Youden’s J .

S6.1 AUROC

Table S12: Per-label AUROC comparison across baseline, contrastive, and autoregressive models on the PadChest dataset. Tail abnormalities are marked with an asterisk (*), and medium-prevalence abnormalities are marked with a dagger (†). For each row, the best value is shown in bold and the second-best value is underlined.

Abnormality	EVA-Base	RAD-DINO	ARK	CheXFound	BiomedCLIP	MedKLIP	BioViL-T	Med-CLIP	Med-AR (ours)	Med-AR-8B (ours)
NSG Tube†	0.9695	0.9685	0.9925	<u>0.9928</u>	0.9291	0.9888	0.9653	0.9899	0.9934	0.9919
Volume Loss*	0.7783	0.7486	0.9117	0.9137	0.6554	0.8210	0.5479	0.9087	0.9226	<u>0.9163</u>
Kyphosis†	0.8535	0.8203	0.8683	<u>0.9009</u>	0.6513	0.8491	0.5987	0.8914	0.8988	0.9029
Alveolar Pattern†	0.8863	0.8728	0.9278	0.9270	0.7708	0.9055	0.8291	0.9219	<u>0.9277</u>	0.9204
Pacemaker*	0.9851	0.9819	<u>0.9967</u>	0.9954	0.9244	0.9944	0.9641	0.9954	0.9969	0.9955
COPD Signs	0.8174	0.8084	0.8324	0.8386	0.6645	0.8168	0.6061	0.8318	0.8436	<u>0.8407</u>
Infiltrates	0.7314	0.7132	<u>0.8069</u>	0.8068	0.6065	0.7658	0.6513	0.7880	0.8125	0.8001
Vertebral Degenerative Changes†	0.7564	0.7351	0.7897	<u>0.8368</u>	0.6221	0.7892	0.5485	0.8247	0.8484	0.8366
Apical Pleural Thickening*	0.7702	0.7448	0.8659	0.8721	0.6104	0.7656	0.5514	0.8787	0.8966	<u>0.8796</u>
Pleural Effusion	0.9254	0.8994	0.9578	<u>0.9572</u>	0.8145	0.9454	0.8444	0.9539	0.9578	0.9535
Atelectasis*	0.8360	0.8197	0.8725	<u>0.8763</u>	0.7359	0.8547	0.7866	0.8705	0.8814	0.8704
Central Venous Catheter Via Jugular Vein*	0.9646	0.9602	0.9945	0.9934	0.9399	0.9912	0.9806	<u>0.9949</u>	0.9954	0.9926
Air Trapping†	0.8014	0.7936	0.8043	<u>0.8191</u>	0.6285	0.8042	0.6393	0.8151	0.8241	0.8132
Calcified Granuloma*	0.7086	0.7017	0.8730	0.8381	0.6146	0.7148	0.5435	0.8600	0.8880	<u>0.8769</u>
Interstitial Pattern	0.8179	0.7892	<u>0.8712</u>	0.8688	0.7008	0.8442	0.7646	0.8634	0.8747	0.8645
Aortic Elongation	0.8321	0.8050	0.8847	0.8922	0.6674	0.8630	0.6239	0.8874	0.8989	<u>0.8931</u>
Laminar Atelectasis†	0.7137	0.6822	<u>0.8700</u>	0.8526	0.5733	0.7761	0.5932	0.8413	0.8701	0.8572
Costophrenic Angle Blunting†	0.7862	0.7547	0.8714	<u>0.8741</u>	0.6535	0.8522	0.6852	0.8679	0.8781	0.8708
Scoliosis	0.7402	0.7174	0.7963	0.8561	0.5989	0.7522	0.5378	0.8378	0.8699	<u>0.8685</u>
Chronic Changes	0.7846	0.7731	0.8035	0.8060	0.6826	0.7803	0.6157	0.8142	0.8252	<u>0.8192</u>
Pneumonia	0.7651	0.7490	0.8577	0.8516	0.6109	0.8106	0.6395	0.8312	<u>0.8548</u>	0.8415
Callus Rib Fracture*	0.6777	0.6750	<u>0.8425</u>	0.7856	0.6002	0.6830	0.5561	0.8313	0.8575	0.8398
Sternotomy*	0.9769	0.9805	0.9964	0.9950	0.8824	0.9947	0.9747	<u>0.9963</u>	<u>0.9963</u>	0.9964
Unchanged	0.6619	0.6527	0.6867	0.6859	0.5843	0.6618	0.6262	<u>0.6977</u>	0.7075	0.6972
Vascular Hilar Enlargement†	0.7038	0.6887	0.7467	0.7538	0.6272	0.7005	0.6111	0.7598	0.7776	<u>0.7731</u>
Bronchiectasis*	0.7490	0.7332	<u>0.8380</u>	0.8268	0.6346	0.7812	0.6471	0.8230	0.8430	0.8330
Fibrotic Band*	0.7388	0.7107	0.8562	0.8536	0.6428	0.7436	0.5516	0.8427	0.8659	<u>0.8597</u>
Cardiomegaly	0.8851	0.8696	0.9214	<u>0.9237</u>	0.6998	0.8918	0.7164	0.9193	0.9265	0.9191
Nodule†	0.6698	0.6720	0.7887	0.7684	0.5785	0.6894	0.4982	0.7661	<u>0.7885</u>	0.7696

S6.2 AUPRC

Table S13: Per-label AUPRC comparison across baseline, contrastive, and autoregressive models on the PadChest dataset. Tail abnormalities are marked with an asterisk (*), and medium-prevalence abnormalities are marked with a dagger (†). For each row, the best value is shown in bold and the second-best value is underlined.

Abnormality	EVA-Base	RAD-DINO	ARK	CheXFound	BiomedCLIP	MedKLIP	BioViL-T	Med-CLIP	Med-AR-2B (ours)	Med-AR-8B (ours)
NSG tube†	0.5976	0.5641	0.8669	<u>0.8691</u>	0.4898	0.8268	0.5539	0.8283	0.8728	0.8630
Volume loss*	0.1298	0.1106	0.3903	0.3851	0.0720	0.1751	0.0347	0.3933	0.4516	<u>0.4227</u>
Kyphosis†	0.3222	0.2504	0.3504	<u>0.4490</u>	0.1112	0.2906	0.0687	0.4186	0.4494	0.4390
Alveolar pattern†	0.4612	0.4173	0.5698	0.5612	0.2338	0.5154	0.3237	0.5623	<u>0.5651</u>	0.5504
Pacemaker*	0.7318	0.7209	0.9142	0.9008	0.6551	0.8997	0.5389	0.9164	0.9388	<u>0.9374</u>
COPD signs	0.5282	0.5107	0.5612	0.5695	0.3509	0.5288	0.2584	0.5638	0.5862	<u>0.5796</u>
Infiltrates	0.1607	0.1481	0.2397	<u>0.2463</u>	0.1010	0.1966	0.1115	0.2163	0.2518	0.2418
Vertebral degenerative changes†	0.1326	0.1187	0.1683	0.2387	0.0766	0.1580	0.0542	0.2163	0.2602	<u>0.2461</u>
Apical pleural thickening*	0.1286	0.1088	0.2909	0.2992	0.0707	0.1220	0.0433	0.3374	0.3643	<u>0.3460</u>
Pleural effusion	0.6255	0.5219	<u>0.7803</u>	0.7796	0.3477	0.7169	0.3728	0.7684	0.7824	0.7681
Atelectasis*	0.1504	0.1126	0.2396	<u>0.2413</u>	0.0818	0.1823	0.0992	0.2231	0.2452	0.2325
Central venous catheter via jugular vein*	0.4037	0.3664	0.8416	<u>0.8446</u>	0.3505	0.8060	0.6551	0.8451	0.8682	<u>0.8642</u>
Air trapping†	0.2203	0.2123	0.2206	0.2319	0.0959	0.2158	0.0815	<u>0.2390</u>	0.2401	0.2277
Calcified granuloma*	0.0720	0.0686	0.3702	0.2902	0.0526	0.0702	0.0344	0.3677	0.4253	<u>0.4036</u>
Interstitial pattern	0.3124	0.2650	<u>0.4232</u>	0.4175	0.1858	0.3593	0.2207	0.4112	0.4253	0.4084
Aortic elongation	0.3602	0.3171	0.4735	<u>0.5059</u>	0.1869	0.4347	0.1389	0.4845	0.5202	0.5030
Laminar atelectasis†	0.1174	0.0934	<u>0.4013</u>	0.3733	0.0620	0.2088	0.0594	0.3539	0.4053	0.3913
Costophrenic angle blunting†	0.2182	0.1906	0.4023	0.4045	0.1314	0.3542	0.1214	0.4025	0.4218	<u>0.4071</u>
Scoliosis	0.1914	0.1621	0.3033	0.4324	0.1031	0.2028	0.0768	0.3974	<u>0.4649</u>	0.4664
Chronic changes	0.2159	0.2001	0.2318	0.2422	0.1521	0.2172	0.0961	<u>0.2660</u>	0.2728	0.2617
Pneumonia	0.2601	0.2248	<u>0.4333</u>	0.4231	0.1193	0.3302	0.1259	0.3825	0.4401	0.4118
Callus rib fracture*	0.0517	0.0532	0.3217	0.1570	0.0418	0.0553	0.0317	0.3159	0.3513	<u>0.3249</u>
Sternotomy*	0.6916	0.6944	0.8915	0.8881	0.5077	0.8628	0.5397	<u>0.8883</u>	0.8814	0.8813
Unchanged	0.2034	0.1964	0.2261	0.2258	0.1666	0.2018	0.1876	0.2339	0.2381	<u>0.2370</u>
Vascular hilar enlargement†	0.1017	0.0934	0.1321	0.1344	0.0772	0.0997	0.0637	0.1457	0.1584	<u>0.1570</u>
Bronchiectasis*	0.0864	0.0700	0.1683	0.1621	0.0551	0.1064	0.0561	0.1581	0.1933	<u>0.1861</u>
Fibrotic band*	0.1344	0.1008	0.3190	0.3071	0.0825	0.1262	0.0542	0.3162	0.3526	<u>0.3417</u>
Cardiomegaly	0.6051	0.5630	0.7120	<u>0.7159</u>	0.3205	0.6235	0.2696	0.7041	0.7238	0.7051
Nodule†	0.0608	0.0615	<u>0.1583</u>	0.1305	0.0457	0.0708	0.0322	0.1435	0.1740	0.1548

S6.3 Sensitivity

Table S14: Per-label **Sensitivity** comparison across baseline, contrastive, and autoregressive models on the PadChest dataset. Tail abnormalities are marked with an asterisk (*), and medium-prevalence abnormalities are marked with a dagger (†). For each row, the best value is shown in bold and the second-best value is underlined.

Abnormality	EVA-Base	RAD-DINO	ARK	CheXFound	BiomedCLIP	MedKLIP	BioViL-T	Med-CLIP	Med-AR-2B (ours)	Med-AR-8B (ours)
NSG tube†	0.9554	0.9705	0.9857	0.9777	0.8375	0.9737	0.9450	0.9721	<u>0.9809</u>	0.9785
Volume loss*	0.6830	0.7692	0.8276	<u>0.8382</u>	0.5040	0.7944	0.6631	0.8130	0.8236	0.8554
Kyphosis†	0.8060	0.7610	0.8419	<u>0.8659</u>	0.6247	0.8052	0.8352	0.8479	0.8142	0.8831
Alveolar pattern†	0.8368	0.8318	0.8677	0.8878	0.7549	0.8066	0.7026	0.8601	0.8570	<u>0.8715</u>
Pacemaker*	0.9643	0.9475	0.9855	0.9788	0.7969	0.9777	0.9375	0.9810	<u>0.9900</u>	0.9922
COPD signs	<u>0.7921</u>	0.7526	0.7898	0.7789	0.5750	0.7676	0.7691	0.7855	0.8083	0.7829
Infiltrates	0.7383	0.7278	0.7925	0.7260	0.5320	0.7137	0.5665	0.7525	0.7586	<u>0.7629</u>
Vertebral degenerative changes†	0.8148	0.8131	0.8064	0.7896	0.6563	0.7519	0.9279	0.7611	<u>0.8399</u>	0.8106
Apical pleural thickening*	0.7497	0.7920	0.7584	0.8180	0.5287	0.7443	<u>0.8212</u>	0.7941	0.8256	0.8061
Pleural effusion	0.8464	0.8619	0.9031	0.9157	0.7558	0.8963	0.8007	<u>0.9060</u>	0.9043	0.9022
Atelectasis*	0.7818	0.7629	0.8171	<u>0.8388</u>	0.7290	0.7696	0.7141	0.8279	0.8428	0.8225
Central venous catheter via jugular vein*	0.9697	0.9818	0.9818	<u>0.9806</u>	0.9126	0.9757	0.9575	0.9745	0.9769	0.9636
Air trapping†	0.7697	0.7853	0.8284	<u>0.8187</u>	0.5862	0.8187	0.8692	0.7623	<u>0.8291</u>	0.7801
Calcified granuloma*	0.6493	0.7537	0.7483	0.7309	0.6292	0.7323	0.9277	0.8086	<u>0.8220</u>	0.7403
Interstitial pattern	0.7568	0.7472	0.8347	0.8126	0.6779	0.7784	0.7085	0.8080	<u>0.8246</u>	0.8226
Aortic elongation	0.8629	0.7800	0.8637	0.8523	0.6888	0.8322	<u>0.8731</u>	0.8570	0.8853	0.8695
Laminar atelectasis†	0.7576	0.7732	0.7837	0.7350	<u>0.7915</u>	0.6855	0.8297	0.7515	0.7454	0.7202
Costophrenic angle blunting†	0.7437	0.6718	0.8138	0.7977	0.6391	0.7793	0.6994	<u>0.8144</u>	0.8195	0.8011
Scoliosis	0.7196	0.6700	0.7202	0.7852	0.6468	0.7422	0.9302	0.7763	0.7751	<u>0.8013</u>
Chronic changes	<u>0.8368</u>	0.8049	0.8203	0.8126	0.6698	0.7753	0.8736	0.7907	0.8110	0.8148
Pneumonia	0.6411	0.7501	<u>0.7679</u>	0.7593	0.5947	0.7097	0.6433	0.7453	0.7949	0.7620
Callus rib fracture*	0.7937	<u>0.7831</u>	0.6973	0.7169	0.7154	0.6657	0.6913	0.7229	0.7410	0.7741
Sternotomy*	0.9297	0.9568	<u>0.9932</u>	0.9851	0.7392	0.9770	0.9284	0.9905	0.9959	0.9919
Unchanged	<u>0.7794</u>	0.7473	<u>0.7717</u>	0.6759	0.5058	0.7436	0.6847	0.7324	0.8160	0.6888
Vascular hilar enlargement†	0.7587	0.7708	0.8339	<u>0.7872</u>	0.6298	0.7189	0.7664	0.7517	0.7664	0.7085
Bronchiectasis*	0.8125	0.7217	0.7932	0.7887	0.5417	0.7887	0.5045	0.7812	<u>0.8021</u>	0.7664
Fibrotic band*	0.6467	0.7870	0.7527	0.6899	0.5388	<u>0.7910</u>	0.3052	0.7527	0.7242	0.7939
Cardiomegaly	0.8370	0.8370	0.8749	0.8693	0.6636	<u>0.8813</u>	0.8462	0.8765	0.8886	0.8698
Nodule†	0.6344	0.6370	<u>0.7506</u>	0.6408	0.5659	0.7080	0.9574	0.7080	0.6977	0.7287

S6.4 Specificity

Table S15: Per-label **Specificity** comparison across baseline, contrastive, and autoregressive models on the PadChest dataset. Tail abnormalities are marked with an asterisk (*), and medium-prevalence abnormalities are marked with a dagger (†). For each row, the best value is shown in bold and the second-best value is underlined.

Abnormality	EVA-Base	RAD-DINO	ARK	CheXFound	BiomedCLIP	MedKLIP	BioViL-T	Med-CLIP	Med-AR-2B (ours)	Med-AR-8B (ours)
NSG tube†	0.9266	0.9147	0.9654	<u>0.9715</u>	0.8855	0.9590	0.9095	0.9643	0.9723	0.9712
Volume loss*	0.7478	0.5952	0.8366	<u>0.8373</u>	0.7262	0.7140	0.4524	<u>0.8477</u>	0.8657	0.8266
Kyphosis†	0.7370	0.7264	0.7439	0.7791	0.6014	0.7316	0.3773	<u>0.7823</u>	0.8284	0.7720
Alveolar pattern†	0.7680	0.7514	0.8327	0.8112	0.6522	<u>0.8336</u>	0.8243	0.8334	0.8459	0.8142
Pacemaker*	0.9373	0.9372	<u>0.9900</u>	0.9850	0.9082	0.9771	0.8835	0.9887	0.9938	0.9878
COPD signs	0.6898	0.7143	0.7166	0.7400	0.6602	0.7105	0.3865	0.7168	0.7214	<u>0.7369</u>
Infiltrates	0.6214	0.6082	0.6908	0.7482	0.6384	0.6853	0.6577	0.6896	<u>0.7314</u>	0.6992
Vertebral degenerative changes†	0.5695	0.5372	0.6248	<u>0.7164</u>	0.5338	0.6851	0.2207	0.7396	0.6981	0.7126
Apical pleural thickening*	0.6601	0.5633	<u>0.8104</u>	0.7763	0.6465	0.6454	0.2872	0.8035	0.8113	0.7992
Pleural effusion	0.8692	0.7963	0.8928	0.8810	0.7498	0.8692	0.7499	0.8804	<u>0.8906</u>	0.8843
Atelectasis*	0.7844	0.7646	<u>0.7928</u>	0.7745	0.6293	0.8116	0.7398	0.7806	0.7843	0.7801
Central venous catheter via jugular vein*	0.9097	0.9005	0.9685	0.9605	0.8592	0.9484	0.9265	0.9757	<u>0.9853</u>	0.9859
Air trapping†	<u>0.6949</u>	0.6716	0.6300	0.6659	0.6041	0.6402	0.3453	0.7156	0.6691	<u>0.6949</u>
Calcified granuloma*	0.6689	0.5392	<u>0.8242</u>	0.7911	0.5450	0.5867	0.1760	0.7435	0.7888	0.8485
Interstitial pattern	0.7328	0.7040	0.7536	0.7728	0.6211	0.7610	0.6995	0.7622	<u>0.7671</u>	0.7574
Aortic elongation	0.6506	0.6792	0.7439	0.7695	0.5561	0.7303	0.3615	<u>0.7630</u>	0.7506	0.7613
Laminar atelectasis†	0.5544	0.5013	0.8038	0.8171	0.3144	0.7209	0.3527	0.7744	<u>0.8374</u>	0.8425
Costophrenic angle blunting†	0.6936	0.7096	0.7802	0.7988	0.5838	0.7758	0.5965	0.7684	0.7866	<u>0.7928</u>
Scoliosis	0.6448	0.6539	0.7253	0.7594	0.5085	0.6290	0.1761	0.7355	0.7900	<u>0.7692</u>
Chronic changes	0.6003	0.6168	0.6457	0.6579	0.6129	0.6549	0.3580	<u>0.6858</u>	0.6883	0.6712
Pneumonia	0.7603	0.6334	<u>0.7875</u>	0.7892	0.5768	0.7626	0.5659	0.7605	0.7628	0.7654
Callus rib fracture*	0.4822	0.4698	0.8172	0.7015	0.4378	0.6113	0.4273	0.7746	<u>0.8047</u>	0.7395
Sternotomy*	0.9319	0.9377	<u>0.9922</u>	0.9937	0.8969	0.9764	0.9405	0.9869	0.9908	0.9921
Unchanged	0.4610	0.4743	<u>0.5004</u>	0.5919	0.6113	0.5004	0.5003	0.5601	0.4853	<u>0.6101</u>
Vascular hilar enlargement†	0.5533	0.5131	0.5366	0.5973	0.5647	0.5851	0.4326	0.6375	<u>0.6523</u>	0.6996
Bronchiectasis*	0.5564	0.6251	0.7413	0.7259	0.6800	0.6408	0.7100	0.7291	0.7152	<u>0.7320</u>
Fibrotic band*	0.6948	0.5159	0.7974	0.8444	0.6823	0.5696	0.7795	0.7605	<u>0.8332</u>	0.7573
Cardiomegaly	0.7785	0.7440	<u>0.8192</u>	0.8276	0.6277	0.7497	0.5137	0.8046	0.8134	0.8160
Nodule†	0.6199	0.6157	0.6903	0.7597	0.5649	0.5750	0.1375	0.7033	<u>0.7415</u>	0.6836

S6.5 Youden’s J

Table S16: Per-label **Youden’s J Score** comparison across baseline, contrastive, and autoregressive models on the PadChest dataset. Tail abnormalities are marked with an asterisk (*), and medium-prevalence abnormalities are marked with a dagger (†). For each row, the best value is shown in bold and the second-best value is underlined.

Abnormality	EVA-Base	RAD-DINO	ARK	CheXFound	BiomedCLIP	MedKLIP	BioViL-T	Med-CLIP	Med-AR-2B (ours)	Med-AR-8B (ours)
NSG tube†	0.8820	0.8852	<u>0.9511</u>	0.9492	0.7230	0.9327	0.8545	0.9364	0.9532	0.9497
Volume loss*	0.4308	0.3644	0.6642	0.6755	0.2302	0.5084	0.1155	0.6607	0.6893	<u>0.6820</u>
Kyphosis†	0.5430	0.4874	0.5858	<u>0.6450</u>	0.2261	0.5368	0.2125	0.6302	0.6426	0.6551
Alveolar pattern†	0.6048	0.5832	<u>0.7004</u>	0.6990	0.4071	0.6402	0.5269	0.6935	0.7029	0.6857
Pacemaker*	0.9016	0.8847	0.9755	0.9638	0.7051	0.9548	0.8210	0.9697	0.9838	<u>0.9800</u>
COPD signs	0.4819	0.4669	0.5064	0.5189	0.2352	0.4781	0.1556	0.5023	0.5297	<u>0.5198</u>
Infiltrates	0.3597	0.3360	<u>0.4833</u>	0.4742	0.1704	0.3990	0.2242	0.4421	0.4900	0.4621
Vertebral degenerative changes†	0.3843	0.3503	<u>0.4312</u>	0.5060	0.1901	0.4370	0.1486	0.5007	0.5380	<u>0.5232</u>
Apical pleural thickening*	0.4098	0.3553	0.5688	0.5943	0.1752	0.3897	0.1084	0.5976	0.6369	<u>0.6053</u>
Pleural effusion	0.7156	0.6582	<u>0.7959</u>	0.7967	0.5056	0.7655	0.5506	0.7864	0.7949	0.7865
Atelectasis*	0.5662	0.5275	0.6099	<u>0.6133</u>	0.3583	0.5812	0.4539	0.6085	0.6271	0.6026
Central venous catheter via jugular vein*	0.8794	0.8823	<u>0.9503</u>	0.9411	0.7718	0.9241	0.8840	0.9502	0.9622	0.9495
Air trapping†	0.4646	0.4569	0.4584	<u>0.4846</u>	0.1903	0.4589	0.2145	0.4779	0.4982	0.4750
Calcified granuloma*	0.3182	0.2929	0.5725	0.5220	0.1742	0.3190	0.1037	0.5521	0.6108	<u>0.5888</u>
Interstitial pattern	0.4896	0.4512	<u>0.5883</u>	0.5854	0.2990	0.5394	0.4080	0.5702	0.5917	0.5800
Aortic elongation	0.5135	0.4592	0.6076	0.6218	0.2449	0.5625	0.2346	0.6200	0.6359	<u>0.6308</u>
Laminar atelectasis†	0.3120	0.2745	0.5875	0.5521	0.1059	0.4064	0.1824	0.5259	<u>0.5828</u>	0.5627
Costophrenic angle blunting†	0.4373	0.3814	0.5940	<u>0.5965</u>	0.2229	0.5551	0.2959	0.5828	0.6061	0.5939
Scoliosis	0.3644	0.3239	0.4455	0.5446	0.1553	0.3712	0.1063	0.5118	<u>0.5651</u>	0.5705
Chronic changes	0.4371	0.4217	0.4660	0.4705	0.2827	0.4302	0.2316	0.4765	0.4993	<u>0.4860</u>
Pneumonia	0.4014	0.3835	<u>0.5554</u>	0.5485	0.1715	0.4723	0.2092	0.5058	0.5577	0.5274
Callus rib fracture*	0.2759	0.2529	<u>0.5145</u>	0.4184	0.1532	0.2770	0.1186	0.4975	0.5457	0.5136
Sternotomy*	0.8616	0.8945	<u>0.9854</u>	0.9788	0.6361	0.9534	0.8689	0.9774	0.9867	0.9840
Unchanged	0.2404	0.2216	0.2721	0.2678	0.1171	0.2440	0.1850	0.2925	0.3013	<u>0.2989</u>
Vascular hilar enlargement†	0.3120	0.2839	0.3705	0.3845	0.1945	0.3040	0.1990	0.3892	0.4187	<u>0.4081</u>
Bronchiectasis*	0.3689	0.3468	0.5345	0.5146	0.2217	0.4295	0.2145	0.5103	<u>0.5173</u>	0.4984
Fibrotic band*	0.3415	0.3029	0.5501	0.5343	0.2211	0.3606	0.0847	0.5132	0.5574	<u>0.5512</u>
Cardiomegaly	0.6155	0.5810	0.6941	<u>0.6969</u>	0.2913	0.6310	0.3599	0.6811	0.7020	0.6858
Nodule†	0.2543	0.2527	0.4409	0.4005	0.1308	0.2830	0.0949	0.4113	<u>0.4392</u>	0.4123