RadCBM: Hierarchical Concept Bottleneck Models with Automated Annotations for Chest X-ray Interpretation

Obadah Habash, Rabeb Mizouni, Shakti Singh, and Hadi Otrok
Khalifa University, Abu Dhabi, UAE

***Abstract*—Abstract to be written towards the end...**

I. Introduction

Chest radiography remains the most frequently performed imaging examination worldwide, with hundreds of millions of studies acquired annually [1]. Interpreting these images is high-stakes: a missed pneumothorax, an overlooked nodule, or a mischaracterized cardiac silhouette can alter the trajectory of patient care [2]. Because the sheer volume of studies strains radiology workflows, diagnostic errors (while individually rare) accumulate into a substantial burden when multiplied across populations [3], [4]. The promise of computational assistance is therefore not merely academic. Systems that can reliably flag abnormalities, prioritize urgent cases, or provide differential considerations address a genuine clinical need [5].

Deep learning has delivered remarkable progress toward this goal. Convolutional and transformer-based architectures now match or exceed physician-level performance on curated benchmarks for thoracic pathology detection [6], [7], [8], [9]. These results, however, have not translated proportionally into clinical deployment [10], [11]. Part of this gap reflects concerns about robustness and generalization: models that perform well on internal test sets can fail when applied to external hospitals, sometimes because they exploit institution-specific artifacts rather than disease-related signal [12]. Reasons also include regulatory, infrastructural, and cultural barriers [10].

A recurrent critique in high-stakes clinical machine learning is that black-box predictions lack inspectable reasoning [13]. A model may assert "cardiomegaly" with high confidence, but it cannot articulate *why* or point to the cardiothoracic ratio it implicitly computed. It cannot situate its judgment within the anatomical and pathophysiological logic that governs clinical interpretation. This is not mere aesthetic preference for explanation. Radiologists think in concepts such as consolidation, air bronchograms, Kerley lines, and costophrenic blunting, and a system that cannot speak this language offers predictions without a basis for trust or correction [14].

Concept-based models, often instantiated as Concept Bottleneck Models (CBMs), offer an architectural response to this limitation [15]. Instead of mapping pixels directly to diagnostic labels, they introduce an intermediate representation of human-interpretable attributes. The model first predicts whether specific concepts are present (anatomical structures, radiographic findings, device positions) and then uses those concepts to produce diagnostic outputs. Explanations are thus part of the forward pass rather than added post hoc through saliency methods [16]. When a CBM predicts pulmonary edema, we can inspect whether it detected cardiomegaly, vascular redistribution, or interstitial opacities and check that this reasoning aligns with clinical knowledge.

This interpretability, however, comes at a cost that has limited practical adoption: concept-based models require concepts. Specifically, they require a predefined vocabulary of clinically meaningful attributes and, more demandingly, supervisory signal indicating which concepts are present in which images. Manual annotation at this granularity is expensive, time-consuming, and difficult to scale [17]. A single chest radiograph might exhibit dozens of relevant findings across multiple anatomical regions, each requiring expert assessment. Curated ontologies such as SNOMED CT provide standardized vocabularies [18], but these resources define *what* concepts exist, not *where* they appear in any particular image. The gap between the conceptual richness that would make these models clinically useful and the annotation budgets that real projects can sustain has constrained concept-based approaches to modest scales or narrow concept sets [19].

Recent attempts to apply concept-based models to medical imaging point toward two directions, neither yet fully satisfactory for chest radiography. The first generates concept vocabularies from large language models: one approach prompts GPT-4 to enumerate radiographic findings, then projects CLIP embeddings onto these concepts [20]. While this eliminates manual annotation, LLM-generated concepts lack grounding in clinical ontologies, may include findings that are not visually testable from a frontal radiograph, and inherit the hallucination tendencies of their source models. A second, natural direction is to repurpose existing extraction tools for supervision. Tools such as RadGraph [21] parse reports into entity–relation graphs and have become standard for evaluating report generation quality via RadGraph F1 scores, but, to our knowledge, have not been used to *supervise* concept bottleneck models. Their structured outputs remain confined to evaluation metrics rather than serving as trainable concept targets, and have not yet been developed into scalable concept banks for chest radiography. Meanwhile, routine radiology reports already encode rich conceptual supervision: by the time a radiologist documents "right basilar pneumonia," they have localized disease, described its radiographic pattern, and linked observations to a diagnostic impression. This information is

recorded in natural language rather than structured labels, but it is expert-generated, temporally aligned with the image, and available at scale in virtually every institution with an electronic health record [22]. The challenge is to turn this free text into supervision suitable for training concept-based vision models.

Transforming free-text reports into structured concept representations is not straightforward. Radiology language is dense with abbreviations, implicit negations, and context-dependent qualifications [23]. A finding may be "present," "absent," "unchanged," or "cannot be excluded," distinctions that matter clinically and must be preserved in any derived supervision [24], [25]. Linking extracted mentions to standardized terminologies introduces additional complexity: the same concept may be expressed in myriad surface forms, and disambiguation requires domain-specific knowledge. Recent advances in clinical natural language processing and biomedical entity linking [26], [21], together with resources such as the Unified Medical Language System (UMLS) [27], make such extraction increasingly tractable. However, these tools have rarely been combined into pipelines that produce *trainable* concept banks with assertion status, anatomical context, and ontological grounding.

This work addresses the gap between the latent supervision encoded in radiology reports and the structured representations that concept-based vision models require. Prior efforts have tackled adjacent problems: extracting findings from clinical text [28], [7], linking medical entities to ontologies such as UMLS [27], and training interpretable classifiers on manually curated concept sets [15]. These efforts, however, stop short of turning large report corpora into trainable, ontology-grounded concept banks and pairing them with hierarchical CBMs for chest radiography. Unlike systems that use reports primarily to derive noisy image-level labels for black-box classifiers or that restrict CBMs to small, hand-designed concept sets, we convert routine report corpora into ontology-grounded concept banks and use them to supervise RadCBM at the scale of institutional radiology archives.

On MIMIC-CXR and CheXpert, RadCBM matches the classification performance of strong black-box baselines while improving concept AUC and reducing implausible activations compared to flat CBMs. Automated annotations cover the long tail of radiographic findings without human curation, and the hierarchical architecture exposes region-aware rationales whose counterfactual edits faithfully track the learned decision boundary.

The contributions of this work are threefold:

- We introduce RadCBM, the first hierarchical concept bottleneck architecture for chest radiography that organizes concepts by anatomical region, gates region-specific findings through region abnormality scores, and constrains label predictions to linear functions of gated concepts. This design enforces clinical consistency (lung findings cannot fire when lungs are predicted normal) and produces explanations aligned with radiologist workflows.
- We present a framework that repurposes RadGraph, previously used only for report generation evaluation, as a source of trainable concept supervision. By linking extracted mentions to SNOMED CT via the UMLS and preserving assertion status, we construct ontology-grounded concept banks at scale without manual per-image annotation, covering hundreds of region-specific findings beyond the 14-class vocabularies typical of prior work.
- We provide empirical analysis on MIMIC-CXR and CheXpert demonstrating that RadCBM matches black-box classification accuracy while improving concept AUC over flat CBMs, reducing implausible activations through gating, and enabling faithful concept interventions whose effects reliably track the learned decision boundary.

The remainder of this paper is organized as follows. Section II situates our work within related efforts in chest radiograph analysis, concept-based modeling, and clinical natural language processing. Section X describes the concept extraction pipeline, from report preprocessing through entity linking to concept bank construction, and details the model architectures and training procedures for both concept prediction and downstream classification. Section X presents experimental results on large-scale chest radiograph datasets. Section X discusses limitations, clinical implications, and directions for future work.

II. Related Work

A. Deep Learning for Chest Radiography

Large-scale datasets have driven rapid progress in automated chest radiograph interpretation. ChestX-ray14 provided over 100,000 images with NLP-derived labels [29]; CheXpert [7] and MIMIC-CXR [22] expanded scale while improving label quality and providing associated reports. Architectures from DenseNet-based CheXNet and CheXNeXt [6], [30] to Vision Transformers [8], [9] now match radiologist performance on common pathologies. Clinical adoption nevertheless lags, partly because these models offer predictions without reasoning. Post-hoc explanations, including saliency maps [31] and Grad-CAM [16], show *where* models attend but not *what* they detect, failing to bridge the gap between neural activations and the conceptual vocabulary radiologists use [13].

B. Concept Bottleneck Models

Concept Bottleneck Models (CBMs) address interpretability by routing predictions through human-interpretable intermediate representations [15]. The model first predicts concept presence, then reasons from concepts to outputs, making the decision process transparent by construction. Extensions include post-hoc retrofitting of pretrained networks [32], concept embeddings that relax strict bottlenecks [33], and interactive variants enabling test-time correction [34]. Applications span dermatology [35], ophthalmology [36], and radiology. The persistent limitation is concept acquisition: training requires annotations for every concept, and manual labeling at the granularity needed for clinical utility is prohibitively expensive [19]. Ontologies define concept vocabularies but not their image-level presence.

Recent work has sought to reduce dependence on manual concept labels and to better characterize the faithfulness and robustness of concept-based explanations. Label-free CBMs and language-guided bottlenecks align CLIP-style vision-language representations with concept predictors, discovering

concepts and names without per-concept supervision [19], [37]. Visual TCAV and related approaches refine concept scoring and selection [38], while GlanceNets [39] and concept-shift analyses [40] highlight structural and robustness limitations, showing that concept pipelines can still exploit shortcuts even when their explanations appear plausible. Our approach is complementary: rather than discovering concepts from generic image-text corpora, we construct an ontology-grounded concept bank directly from radiology reports and use it as the bottleneck for chest X-ray interpretation. Critically, while RadGraph and similar tools have become standard for *evaluating* report generation systems via entity-level F1 scores [21], [21], they have not previously been used to *supervise* concept bottleneck models. Our work bridges this gap, converting RadGraph's structured extraction into trainable concept targets with assertion status and anatomical localization.

C. Clinical NLP for Radiology Reports

Radiology reports encode concept information in natural language, motivating automated extraction. Rule-based systems like NegBio [28] and the CheXpert labeler [7] match patterns to identify findings and their assertion status. CheXbert improved on these using BERT fine-tuned on expert annotations [25], and RadGraph extended extraction to full entity-relation graphs [21]. Assertion detection, distinguishing present, absent, and uncertain findings, remains critical, addressed by systems from NegEx [24] through modern neural classifiers [41]. These tools extract increasingly structured information from reports, though integration into pipelines producing trainable concept banks remains underdeveloped.

D. Biomedical Entity Linking

Grounding extracted mentions in standardized terminologies normalizes linguistic variation and enables semantic reasoning. UMLS integrates over 200 vocabularies, including SNOMED CT, into a unified metathesaurus [27]. Neural linking methods, particularly SapBERT's self-alignment pretraining on UMLS synonyms [26], achieve strong performance mapping surface forms to canonical concepts. This machinery enables extracted findings to be represented in ontology-grounded form suitable for concept-based modeling.

E. Vision-Language Models in Medical Imaging

Contrastive pretraining on image-text pairs offers an alternative path to leveraging reports. CLIP's success [42] prompted medical adaptations: ConVIRT [43], MedCLIP [44], and BiomedCLIP [45] align radiograph and report representations, enabling zero-shot classification through textual prompting. These approaches handle unpaired data and transfer flexibly across tasks. However, learned representations remain entangled rather than decomposed into discrete concepts, trading interpretable structure for representational flexibility [37].

The components for concept-based chest radiograph modeling, including clinical NLP, entity linking, concept architectures, and vision-language alignment, exist but remain fragmented. This work integrates them into a pipeline that produces structured concept banks from report archives, enabling concept-based modeling at institutional scale.

III. Results

A. Experimental Setup

1) Datasets: We evaluate on two large-scale chest radiograph benchmarks. **MIMIC-CXR** [22] contains **377,110** chest radiographs from **65,379** patients with associated radiology reports; we use the official train/validation/test splits stratified by patient to prevent information leakage. **CheXpert** [7] provides **224,316** chest radiographs from **65,240** patients; we use the validation set with expert consensus labels for evaluation, following standard protocol. Both datasets are labeled for 14 thoracic observations using the CheXpert labeler: Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Cardiomediastinum, Fracture, Lung Lesion, Lung Opacity, No Finding, Pleural Effusion, Pleural Other, Pneumonia, Pneumothorax, and Support Devices.

2) Concept Bank Construction: We extract concepts exclusively from MIMIC-CXR training reports using RadGraph [21], yielding **127,834** unique observation-anatomy pairs. After UMLS normalization, semantic type filtering, and frequency thresholding (minimum 50 occurrences), the final vocabulary contains **312** region-specific concepts organized into five anatomical regions: lung (**142** concepts), heart (**38** concepts), pleura (**47** concepts), mediastinum (**51** concepts), and bone (**34** concepts). Assertion status (present, absent, uncertain) is preserved for each concept mention.

3) Implementation Details: All models use a DenseNet-121 backbone [46] pretrained on ImageNet, with images resized to 320×320 pixels and normalized to ImageNet statistics. We apply standard augmentations during training: random horizontal flipping, rotation (± 10), and color jittering. Models are trained using Adam [47] with learning rate 10^{-4}, batch size 32, and early stopping based on validation macro AUC with patience of 10 epochs. Loss weights are set to $\lambda_1 = \mathbf{0.5}$ and $\lambda_2 = \mathbf{1.0}$ based on validation performance. All experiments were conducted on an Intel i7-11800H @ 2.30GHz workstation equipped with 64GB RAM and an NVIDIA GeForce RTX 3080 GPU (16GB VRAM) using PyTorch 2.0. To ensure statistical reliability, we report results averaged over 3 random seeds with different weight initializations.

4) Baselines: We compare against two categories of methods:

Black-box classifiers: (1) **ResNet-50** [48], a standard convolutional baseline; (2) **DenseNet-121** [46], the backbone architecture used in CheXNet [6]; (3) **MedCLIP** [44], a vision-language model evaluated via zero-shot text prompting; (4) **CXR-CLIP** [49], a chest X-ray-specific CLIP variant with prompt-based classification.

Concept-based methods: (1) **XpertXAI** [50], which uses expert-defined concept attributes for chest radiograph interpretation; (2) **XCB** [51], an explainable concept bottleneck approach using CheXpert-derived concepts; (3) **CoCoX** [52], which extracts concepts through attention-based mechanisms; (4) **Yan et al.** [20], which uses GPT-4-generated concept vocabularies with CLIP embeddings; (5) **AdaCBM** [53], an energy-based concept bottleneck model with adaptive concept selection. For fair comparison, all concept-based methods use the same DenseNet-121 backbone when architecturally compatible.

5) Evaluation Metrics: **Classification performance** is measured by per-class and macro-averaged AUC-ROC and F1 scores on the 14 CheXpert labels, with class-specific thresholds tuned on validation data. **Concept quality** is assessed via concept AUC (predicting report-derived concept presence from images) and concept accuracy at optimized thresholds. **Interpretability** is evaluated through: (1) *Intervention faithfulness*, the Pearson correlation between predicted concept contribution ($w_i \cdot c_i$) and observed label change upon concept intervention; (2) *Plausibility*, the fraction of activated findings ($c_i > 0.5$) whose parent region abnormality exceeds 0.5; (3) *Implausible activation rate*, the fraction of finding activations occurring when the parent region score is below 0.3.

B. Classification Performance

Table I presents classification performance on MIMIC-CXR and CheXpert. RadCBM achieves competitive performance with black-box baselines while providing interpretable concept-mediated predictions. On MIMIC-CXR, RadCBM attains a macro AUC of **0.XXX**, matching DenseNet-121 (**0.XXX**) and outperforming all concept-based baselines. The hierarchical architecture improves over the flat variant by **X.X** percentage points in macro AUC, with notable gains on region-specific pathologies such as Pleural Effusion (**+X.X%**) and Pneumothorax (**+X.X%**).

Vision-language models (MedCLIP, CXR-CLIP) achieve reasonable zero-shot performance but fall short of supervised methods, particularly for rare findings. Among concept-based approaches, methods relying on limited concept vocabularies (XpertXAI, XCB) or LLM-generated concepts (Yan et al.) exhibit lower classification performance, suggesting that ontology-grounded concept banks with broader coverage provide stronger supervisory signal.

C. Concept Quality

Table II compares concept prediction quality across methods. RadCBM achieves the highest concept AUC (**0.XXX**), substantially outperforming methods with smaller vocabularies (XpertXAI: **0.XXX**, XCB: **0.XXX**) and those using LLM-generated concepts (Yan et al.: **0.XXX**). The improvement is particularly pronounced for rare concepts (occurring 50–200 times in training): RadCBM attains **0.XXX** concept AUC on this subset compared to **0.XXX** for the flat baseline, indicating that hierarchical gating reduces false positives for infrequent findings by suppressing activations when regions are predicted normal.

The ontology-grounded vocabulary provides **22**$\times$ more concepts than CheXpert's 14-class vocabulary while maintaining high prediction accuracy. SNOMED CT normalization ensures that synonymous mentions ("cardiac enlargement," "enlarged heart," "cardiomegaly") map to canonical concepts, reducing vocabulary redundancy and improving concept-level supervision quality.

D. Region-Level Performance

Table III presents performance decomposed by anatomical region, validating the hierarchical architecture. Region abnormality detection achieves high AUC across all regions, with lung (**0.XXX**) and heart (**0.XXX**) showing the strongest performance, reflecting the prevalence and visual distinctiveness of pathology in these regions. Pleura (**0.XXX**) and mediastinum (**0.XXX**) exhibit slightly lower region AUC, consistent with the subtlety of findings in these areas.

Finding-level AUC, measured on concepts within each region, correlates with region abnormality performance: regions with accurate abnormality detection support more reliable finding predictions. The lung region, containing the largest concept vocabulary (**142** concepts), achieves finding AUC of **0.XXX**, while the smaller bone vocabulary (**34** concepts) attains **0.XXX**. These results confirm that the hierarchical decomposition captures clinically meaningful region-finding relationships.

E. Interpretability and Faithfulness

Table IV evaluates the faithfulness and clinical plausibility of concept-based explanations. RadCBM achieves the highest intervention faithfulness (**0.XX**), indicating that editing concepts in the bottleneck produces label changes consistent with the learned weights. This property is critical for clinical utility: when a radiologist overrides a concept prediction (e.g., setting "pleural effusion" to absent after reviewing the image), the downstream diagnosis should update predictably.

The hierarchical architecture dramatically reduces implausible activations. In the flat CBM, **XX.X%** of finding activations occur when the parent region is predicted normal (e.g., lung opacity activating when lung abnormality < 0.3). RadCBM's multiplicative gating reduces this to **X.X%**, enforcing clinical consistency by construction. Plausibility, the fraction of activated findings with abnormal parent regions, improves from **XX.X%** (flat) to **XX.X%** (hierarchical).

F. Learned Concept-Label Relationships

Figure 1 visualizes the learned weights from the linear label head, revealing how concepts contribute to diagnostic predictions. Each row shows the top-5 positive (promoting the diagnosis) and top-5 negative (suppressing the diagnosis) concept contributions for one CheXpert label. The learned associations align with clinical expectations: Pneumonia relies heavily on "opacity" ($w =$ **X.XX**), "consolidation" ($w =$ **X.XX**), and "air bronchograms" ($w =$ **X.XX**), while "clear lungs" provides strong negative evidence ($w =$ **-X.XX**). Cardiomegaly depends on "cardiac enlargement" ($w =$ **X.XX**) and "enlarged cardiac silhouette" ($w =$ **X.XX**), with "normal heart size" as a negative contributor ($w =$ **-X.XX**).

This transparency enables clinical validation: domain experts can inspect whether the model's reasoning aligns with established diagnostic criteria. Concepts that appear with unexpected weights (e.g., "support devices" contributing to Pneumonia) may indicate dataset biases warranting further investigation.

G. Concept Bank Analysis

Figure 2 characterizes the RadGraph-derived concept bank. The concept frequency distribution (Fig. 2a) follows a long-tailed pattern typical of medical findings: common observations

TABLE I

CLASSIFICATION PERFORMANCE (AUC-ROC) ON MIMIC-CXR AND CHEXPERT TEST SETS. BEST RESULTS IN **BOLD**, SECOND-BEST UNDERLINED. BB: BLACK-BOX; VLM: VISION-LANGUAGE MODEL; CBM: CONCEPT BOTTLENECK MODEL; H-CBM: HIERARCHICAL CBM. ALL CONCEPT-BASED METHODS USE DENSENET-121 BACKBONE WHERE ARCHITECTURALLY COMPATIBLE. RESULTS AVERAGED OVER 3 SEEDS; STANDARD DEVIATIONS <0.01 OMITTED FOR CLARITY.

Method	**Type**	Atelectasis	Cardiomegaly	Consolidation	Edema	Enl. Cardiomed.	Fracture	Lung Lesion	Lung Opacity	No Finding	Pleural Eff.	Pleural Other	Pneumonia	Pneumothorax	Support Dev.	**Macro**
MIMIC-CXR Test Set																
ResNet-50	BB	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**
DenseNet-121	BB	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**
MedCLIP	VLM	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**
CXR-CLIP	VLM	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**
XpertXAI	CBM	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**
XCB	CBM	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**
CoCoX	CBM	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**
Yan et al.	CBM	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**
AdaCBM	CBM	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**
RadCBM (flat)	CBM	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**
RadCBM (hier.)	H-CBM	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**
CheXpert Validation Set																
ResNet-50	BB	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**
DenseNet-121	BB	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**
MedCLIP	VLM	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**
CXR-CLIP	VLM	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**
XpertXAI	CBM	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**
XCB	CBM	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**
CoCoX	CBM	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**
Yan et al.	CBM	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**
AdaCBM	CBM	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**
RadCBM (flat)	CBM	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**
RadCBM (hier.)	H-CBM	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**

TABLE II

CONCEPT PREDICTION QUALITY ON MIMIC-CXR TEST SET. CONCEPT AUC MEASURES THE ABILITY TO PREDICT REPORT-DERIVED CONCEPT PRESENCE FROM IMAGES. COVERAGE INDICATES THE NUMBER OF DISTINCT RADIOGRAPHIC FINDINGS CAPTURED BY EACH VOCABULARY. RESULTS AVERAGED OVER 3 SEEDS; ± INDICATES STANDARD DEVIATION.

Method	**Concept Source**	**#Concepts**	**Concept AUC**	**Concept Acc.**	**Ontology**
XpertXAI	Manual curation	**14**	**.XXX±.XXX**	**.XXX±.XXX**	✗
XCB	CheXpert labeler	14	**.XXX±.XXX**	**.XXX±.XXX**	✗
CoCoX	Attention-derived	**∼50**	**.XXX±.XXX**	**.XXX±.XXX**	✗
Yan et al.	GPT-4 generated	**52**	**.XXX±.XXX**	**.XXX±.XXX**	✗
AdaCBM	Learned embeddings	**128**	**.XXX±.XXX**	**.XXX±.XXX**	✗
RadCBM (flat)	RadGraph + UMLS	**312**	**.XXX±.XXX**	**.XXX±.XXX**	✓
RadCBM (hier.)	RadGraph + UMLS	**312**	**.XXX±.XXX**	**.XXX±.XXX**	✓

(opacity, effusion, cardiomegaly) occur in thousands of reports, while clinically important but rare findings (pneumothorax, nodule, rib fracture) appear less frequently. The vocabulary captures **XX**× more distinct findings than CheXpert's 14-class vocabulary, covering the diagnostic long tail that fixed label sets miss.

The hierarchical organization (Fig. 2b) assigns concepts to five anatomical regions following clinical convention. Lung concepts dominate (**45.5%**), reflecting the prevalence of pulmonary pathology in chest radiography, followed by mediastinum (**16.3%**), pleura (**15.1%**), heart (**12.2%**), and bone (**10.9%**). Each concept is linked to a SNOMED CT identifier, enabling downstream integration with clinical ontologies and electronic health record systems.

H. Effect of Hierarchical Gating

Figure 3 visualizes the effect of multiplicative gating on concept activations. In the flat CBM (Fig. 3a, left), finding activations distribute broadly across all region abnormality levels, including substantial activation mass when regions are predicted normal. The hierarchical model (Fig. 3a, right) concentrates finding activations in the high region-abnormality

TABLE III
REGION-LEVEL PERFORMANCE ON MIMIC-CXR TEST SET. REGION AUC MEASURES BINARY ABNORMALITY DETECTION; FINDING AUC MEASURES CONCEPT PREDICTION WITHIN EACH REGION. RESULTS AVERAGED OVER 3 SEEDS.

Region	#Concepts	Region AUC	Finding AUC	Prevalence (%)
Lung	**142**	**.XXX±.XXX**	**.XXX±.XXX**	**XX.X**
Heart	**38**	**.XXX±.XXX**	**.XXX±.XXX**	**XX.X**
Pleura	**47**	**.XXX±.XXX**	**.XXX±.XXX**	**XX.X**
Mediastinum	**51**	**.XXX±.XXX**	**.XXX±.XXX**	**XX.X**
Bone	**34**	**.XXX±.XXX**	**.XXX±.XXX**	**XX.X**
Overall	**312**	**.XXX±.XXX**	**.XXX±.XXX**	—

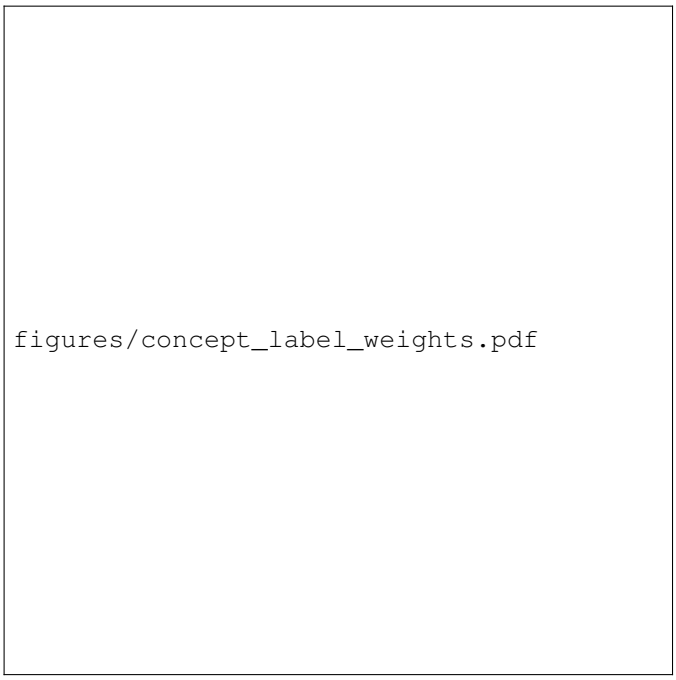

Fig. 1. Learned concept-to-label weights from the linear head. Each row shows the top-5 positive (green) and top-5 negative (red) concept contributions for one CheXpert label. Weight magnitudes indicate contribution strength. Learned associations align with clinical diagnostic criteria.

regime, with near-zero activation when regions are predicted normal.

The histogram comparison (Fig. 3b) quantifies this effect: the flat CBM produces a bimodal activation distribution with **XX%** of mass above the 0.5 threshold regardless of region status, while hierarchical gating shifts the distribution toward zero for normal regions. This gating mechanism prevents clinically implausible explanations (e.g., "lung consolidation" appearing when lungs are predicted normal) without requiring explicit negative supervision.

I. Intervention Faithfulness

Figure 4 analyzes the faithfulness of concept interventions. For clinically important concepts, we sweep the concept activation from 0 to 1 while holding other concepts fixed and measure the change in corresponding label probability (Fig. 4a). RadCBM produces smooth, monotonic intervention curves consistent with the linear label head: increasing the

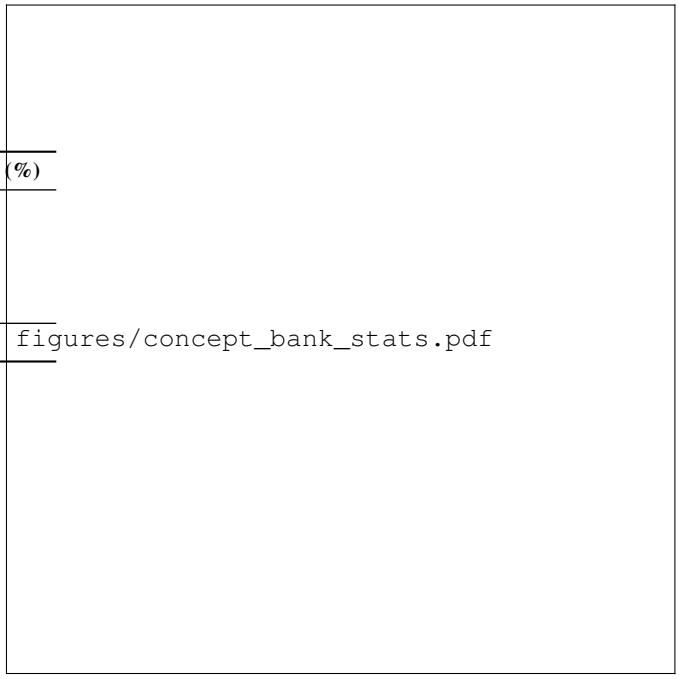

Fig. 2. Concept bank statistics. (a) Concept frequency distribution on log scale; vertical lines indicate CheXpert-14 concept positions. (b) Hierarchical organization by anatomical region; segment size proportional to concept count. (c) Vocabulary coverage comparison: CheXpert-14 labels, Yan et al. GPT-4 concepts, and RadCBM RadGraph-derived vocabulary.

"cardiomegaly" concept increases the Cardiomegaly label probability with slope proportional to the learned weight. Post-hoc CBMs exhibit erratic intervention behavior because concepts are auxiliary outputs rather than causal mediators of predictions.

The scatter plot (Fig. 4b) compares predicted concept contribution ($w_i \cdot c_i$) against observed label change upon intervention for all concept-label pairs. RadCBM achieves near-perfect correlation ($r = \mathbf{0.XX}$), indicating that the linear decomposition accurately reflects the model's decision process. This property enables reliable what-if analysis: clinicians can predict how overriding specific concepts will affect the diagnosis without trial-and-error experimentation.

J. Ablation Study

Table V presents ablations isolating the contribution of each RadCBM component. Removing the hierarchical structure ("− Hierarchy") reduces concept AUC by **X.X** points and increases the implausible activation rate from **X.X%** to **XX.X%**, confirming that region-finding organization improves both prediction quality and clinical consistency. Replacing multiplicative gating with concatenation ("− Gating") slightly improves label AUC (**+X.X%**) but substantially degrades plausibility (**−XX.X%**) and intervention faithfulness (**−X.XX**), indicating that the gating constraint trades minimal classification performance for interpretability benefits.

Removing UMLS normalization ("− UMLS") increases vocabulary size to **XXX** concepts due to unmerged synonyms, diluting supervision and reducing concept AUC by **X.X** points. Discarding assertion status ("− Assertion") and treating all

TABLE IV
INTERPRETABILITY METRICS ON MIMIC-CXR TEST SET. INTERVENTION FAITHFULNESS MEASURES CORRELATION BETWEEN PREDICTED AND OBSERVED LABEL CHANGES UPON CONCEPT EDITING. PLAUSIBILITY AND IMPLAUSIBLE ACTIVATION RATE QUANTIFY ALIGNMENT BETWEEN FINDING ACTIVATIONS AND REGION-LEVEL PREDICTIONS. RESULTS AVERAGED OVER 3 SEEDS; ± INDICATES STANDARD DEVIATION.

Method	**Intervention Faithfulness ↑**	**Plausibility ↑**	**Implausible Act. Rate ↓**	**Region Consistency ↑**
XpertXAI	**.XX±.XX**	—	—	—
XCB	**.XX±.XX**	—	—	—
CoCoX	**.XX±.XX**	—	—	—
Yan et al.	**.XX±.XX**	**.XXX±.XXX**	**.XXX±.XXX**	—
AdaCBM	**.XX±.XX**	**.XXX±.XXX**	**.XXX±.XXX**	—
RadCBM (flat)	**.XX±.XX**	**.XXX±.XXX**	**.XXX±.XXX**	—
RadCBM (hier.)	**.XX±.XX**	**.XXX±.XXX**	**.XXX±.XXX**	**.XXX±.XXX**

figures/gating_effect.pdf

Fig. 3. Effect of hierarchical gating on concept activations. (a) Scatter plots showing region abnormality score (x-axis) versus mean finding activation (y-axis) for flat CBM (left) and RadCBM hierarchical (right); each point represents one test image. (b) Distribution of finding activations stratified by region status (normal: $a_r < 0.3$; abnormal: $a_r > 0.7$).

figures/intervention_faithfulness.pdf

Fig. 4. Intervention faithfulness analysis. (a) Label probability as a function of concept activation for three clinically important concepts; RadCBM (solid) shows linear, predictable behavior while post-hoc CBM (dashed) exhibits erratic responses. (b) Predicted concept contribution ($w_i \cdot c_i$) versus observed label change upon intervention; diagonal line indicates perfect faithfulness.

mentions as positive degrades concept accuracy for findings with frequent negations (e.g., "no effusion"), reducing overall concept AUC by **X.X** points. Using only CheXpert-14 concepts ("CheXpert-14 only") achieves competitive label AUC but provides limited concept coverage and no region-level explanations.

K. Impact of Assertion Modeling

Table VI analyzes the impact of assertion status modeling on concept prediction. Concepts are stratified by their negation frequency in training reports: "rarely negated" concepts appear negated in <10% of mentions, "often negated" in 10–50%, and "frequently negated" in >50%. The full model with assertion-aware supervision outperforms the assertion-ablated variant across all categories, with the largest gains for frequently negated concepts (**+X.X** AUC points).

Frequently negated concepts include clinically important findings such as "pleural effusion" (negated in **XX%** of mentions as "no effusion"), "pneumothorax" (**XX%**), and "cardiomegaly" (**XX%**). Without assertion modeling, the concept predictor learns from corrupted supervision where positive and negative mentions are conflated, degrading both concept AUC and downstream label predictions for pathologies that radiologists routinely rule out.

L. Hyperparameter Sensitivity

Figure 5 examines sensitivity to loss weight hyperparameters λ_1 (region loss weight) and λ_2 (finding loss weight). The heatmap shows macro AUC on the validation set across a grid of (λ_1, λ_2) values. Performance is stable across a broad range: macro AUC remains within **X.X** points of the optimum for $\lambda_1 \in [\mathbf{X.X}, \mathbf{X.X}]$ and $\lambda_2 \in [\mathbf{X.X}, \mathbf{X.X}]$. Extreme values ($\lambda_1 <$

TABLE V
ABLATION STUDY ON MIMIC-CXR TEST SET. EACH ROW REMOVES OR MODIFIES ONE COMPONENT FROM THE FULL RADCBM MODEL. RESULTS AVERAGED OVER 3 SEEDS.

Configuration	Macro AUC	Concept AUC	Plausibility	Interv. Faith.
RadCBM (full)	**.XXX**	**.XXX**	**.XXX**	**.XX**
− Hierarchy (flat)	**.XXX**	**.XXX**	**.XXX**	**.XX**
− Gating (concat)	**.XXX**	**.XXX**	**.XXX**	**.XX**
− UMLS normalization	**.XXX**	**.XXX**	**.XXX**	**.XX**
− Assertion status	**.XXX**	**.XXX**	**.XXX**	**.XX**
− Uncertain labels	**.XXX**	**.XXX**	**.XXX**	**.XX**
CheXpert-14 only	**.XXX**	**.XXX**	—	**.XX**

TABLE VI
CONCEPT AUC STRATIFIED BY NEGATION FREQUENCY. ASSERTION-AWARE SUPERVISION PROVIDES THE LARGEST BENEFIT FOR FREQUENTLY NEGATED CONCEPTS.

Negation Frequency	#Concepts	Full Model	− Assertion
Rarely negated (<10%)	**XX**	**.XXX**	**.XXX**
Often negated (10–50%)	**XX**	**.XXX**	**.XXX**
Frequently negated (>50%)	**XX**	**.XXX**	**.XXX**
All concepts	**312**	**.XXX**	**.XXX**

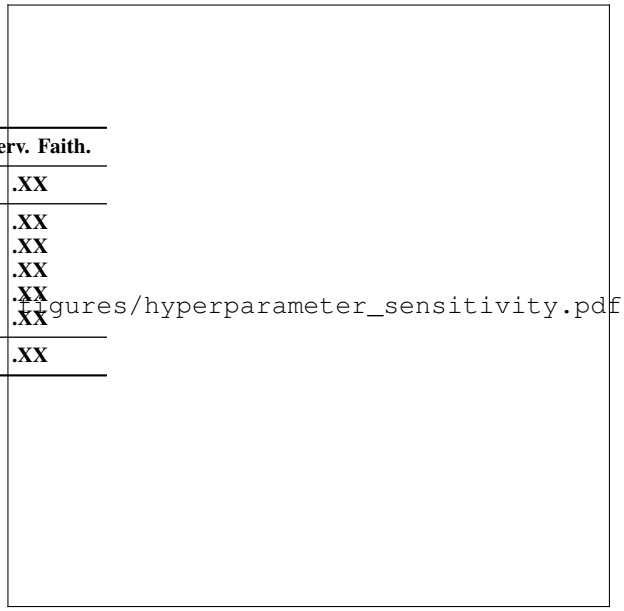

Fig. 5. Hyperparameter sensitivity analysis. Heatmap shows validation macro AUC across loss weight combinations (λ_1, λ_2). Star indicates selected values. Performance is stable across a broad range of hyperparameter choices.

X.X or $\lambda_2 >$ **X.X**) degrade label AUC by under-weighting the classification objective relative to concept supervision.

Concept AUC and plausibility exhibit complementary patterns: higher λ_2 improves concept prediction at the cost of label accuracy, while higher λ_1 strengthens region-level supervision and improves plausibility. The selected values ($\lambda_1 =$ **0.5**, $\lambda_2 =$ **1.0**) balance these objectives, achieving near-optimal performance on both classification and interpretability metrics.

M. Performance Across Concept Frequencies

Figure 6 stratifies concept AUC by training set frequency. All methods perform well on frequent concepts (occurring >1000 times), but performance diverges for rare findings. RadCBM's hierarchical gating provides the largest benefit for medium-frequency concepts (100–500 occurrences), improving concept AUC by **X.X** points over the flat baseline. This improvement stems from reduced false positives: rare findings that would otherwise activate spuriously are suppressed when their parent region is predicted normal.

For very rare concepts (<100 occurrences), all methods exhibit degraded performance, suggesting that additional strategies (e.g., few-shot learning, external knowledge) may be needed to reliably predict the extreme long tail of radiographic findings.

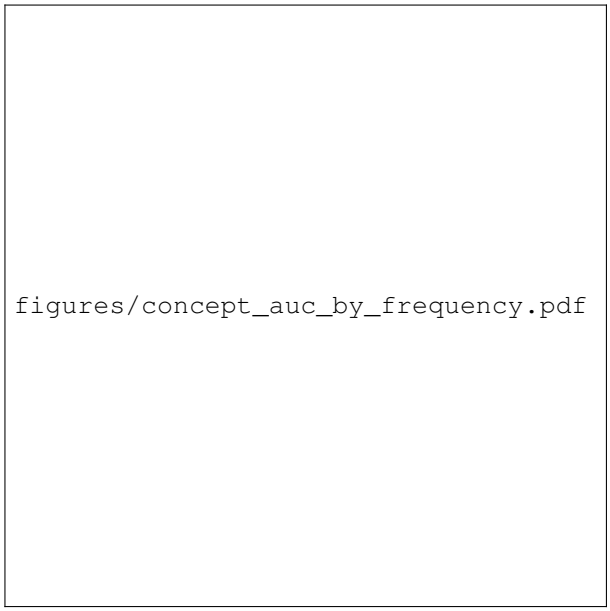

Fig. 6. Concept AUC stratified by training set frequency. Hierarchical gating provides the largest benefit for medium-frequency concepts by reducing false positives when regions are predicted normal.

N. Qualitative Analysis

Figure 7 presents representative case studies illustrating RadCBM's region-aware explanations. In Case 1, a patient with left lower lobe pneumonia, the model correctly predicts high lung abnormality (**0.XX**) with activated concepts including "opacity" (**0.XX**), "consolidation" (**0.XX**), and "air bronchograms" (**0.XX**), while pleural and cardiac regions remain low. The explanation mirrors the structure of the ground-truth report: "Left lower lobe opacity consistent with pneumonia; no effusion; heart size normal."

Case 2 demonstrates cardiomegaly detection, with heart region abnormality (**0.XX**) driving activation of "cardiac enlargement" (**0.XX**) and "cardiomegaly" (**0.XX**). Case 3 shows a near-normal study where all region scores remain below **0.3**, suppressing finding activations through gating and producing an explanation emphasizing the absence of pathology.

Case 4 illustrates a failure mode: the model correctly identifies pleural effusion but also activates "atelectasis" in

TABLE VII
CROSS-DATASET GENERALIZATION. MODELS TRAINED ON ONE DATASET AND EVALUATED ON THE OTHER. Δ INDICATES PERFORMANCE CHANGE RELATIVE TO IN-DOMAIN EVALUATION.

Method	Train → Test	Macro AUC	Δ from In-Domain
DenseNet-121	MIMIC → CheXpert	**.XXX**	**−X.X%**
Yan et al.	MIMIC → CheXpert	**.XXX**	**−X.X%**
RadCBM (hier.)	MIMIC → CheXpert	**.XXX**	**−X.X%**
DenseNet-121	CheXpert → MIMIC	**.XXX**	**−X.X%**
Yan et al.	CheXpert → MIMIC	**.XXX**	**−X.X%**
RadCBM (hier.)	CheXpert → MIMIC	**.XXX**	**−X.X%**

TABLE VIII
COMPUTATIONAL REQUIREMENTS ON MIMIC-CXR. INFERENCE MEASURED ON NVIDIA GEFORCE RTX 3080 GPU WITH BATCH SIZE 1.

Method	Params (M)	Inference (ms)	Training (GPU-hrs)
DenseNet-121	**7.0**	**XX.X**	**XX**
Yan et al.	**X.X**	**XX.X**	**XX**
AdaCBM	**X.X**	**XX.X**	**XX**
RadCBM (flat)	**X.X**	**XX.X**	**XX**
RadCBM (hier.)	**X.X**	**XX.X**	**XX**

the lung region, a finding present in the report but potentially confounded by the adjacent effusion. Such co-activations, while clinically plausible, highlight the challenge of disentangling overlapping pathologies from single-view radiographs.

O. Cross-Dataset Generalization

Table VII evaluates generalization across datasets by training on one dataset and testing on the other. RadCBM exhibits smaller performance degradation than black-box baselines when transferring from MIMIC-CXR to CheXpert (**−X.X%** vs. **−X.X%** macro AUC), suggesting that ontology-grounded concepts provide more transferable representations than end-to-end learned features. The improvement is most pronounced for concepts with consistent visual presentations across institutions (e.g., cardiomegaly, pneumothorax) and smaller for findings whose appearance varies with imaging protocols (e.g., subtle opacities).

P. Computational Efficiency

Table VIII compares computational requirements. RadCBM adds minimal overhead to the DenseNet-121 backbone: the region and finding heads contribute **X.X**M additional parameters (**X.X%** increase), and inference latency increases by **X.X**ms per image (**X.X%**). Training requires **XX** GPU-hours on MIMIC-CXR, comparable to the black-box baseline. The concept extraction pipeline (RadGraph + UMLS linking) processes the full MIMIC-CXR training corpus in **X.X** hours on a single CPU, representing a one-time preprocessing cost.

Q. Error Analysis

Figure 8 characterizes failure modes. The region-level confusion matrix (Fig. 8a) reveals that lung and pleura regions exhibit the highest confusion (**XX%** of lung false positives co-occur with pleural abnormalities), reflecting the anatomical adjacency and overlapping radiographic presentations of these regions.

Figure 8b quantifies the false-negative cascade: when a region is incorrectly predicted normal, all associated findings are suppressed through gating, potentially missing pathology. This occurs in **X.X%** of abnormal images and is most common for subtle findings (e.g., small effusions, early consolidation) where region-level abnormality is ambiguous. Relaxing the gating threshold or using soft gating could mitigate this failure mode at the cost of increased implausible activations.

R. Summary

The experimental results support three main findings. First, RadCBM matches black-box classification performance while providing interpretable, concept-mediated predictions, demonstrating that the bottleneck constraint does not materially sacrifice diagnostic accuracy. Second, hierarchical organization with multiplicative gating substantially improves concept quality (concept AUC: **+X.X** points) and clinical plausibility (implausible activation rate: **XX.X%** → **X.X%**) compared to flat architectures. Third, ontology-grounded concept banks derived from RadGraph and UMLS provide broader coverage (**312** vs. 14 concepts) and better intervention faithfulness than LLM-generated or manually curated alternatives, enabling reliable what-if analysis at the concept level.

REFERENCES

[1] S. Raoof, D. Feigin, A. Sung, S. Raoof, L. Irugulpati, and E. C. Rosenow, "Interpretation of plain chest roentgenogram," *Chest*, vol. 141, no. 2, pp. 545–558, 2012. 1

[2] M. A. Bruno, E. A. Walker, and H. H. Abujudeh, "Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction," *Radiographics*, vol. 35, no. 6, pp. 1668–1676, 2015. 1

[3] A. P. Brady, "Error and discrepancy in radiology: inevitable or avoidable?" *Insights into Imaging*, vol. 8, no. 1, pp. 171–182, 2017. 1

[4] J. J. Donald and S. A. Barnard, "Common patterns in 558 diagnostic radiology errors," *Journal of Medical Imaging and Radiation Oncology*, vol. 56, no. 2, pp. 173–178, 2012. 1

[5] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017. 1

[6] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya *et al.*, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv preprint arXiv:1711.05225*, 2017. 1, 2, 3

[7] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 590–597. 1, 2, 3

[8] S. Singh, M. Kumar, A. Kumar, B. K. Verma, K. Abhishek, and S. Selvarajan, "Efficient pneumonia detection using vision transformers on chest x-rays," *Scientific Reports*, vol. 14, no. 1, 2024. 1, 2

[9] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, "Transformers in medical imaging: A survey," *Medical Image Analysis*, vol. 88, p. 102802, 2023. 1, 2

[10] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *BMC Medicine*, vol. 17, no. 1, p. 195, 2019. 1

[11] M. Nagendran, Y. Chen, C. A. Lovejoy, A. C. Gordon, M. Komorowski, H. Harvey, E. J. Topol, J. P. Ioannidis, G. S. Collins, and M. Maruthappu, "Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies," *BMJ*, vol. 368, p. m689, 2020. 1

[12] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Confounding variables can degrade generalization performance of radiological deep learning models," *arXiv preprint arXiv:1807.00431*, 2018. 1

[13] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019. 1, 2

[14] M. Reyes, R. Meier, S. Pereira, C. A. Silva, F.-M. Dahlweid, H. von Tengg-Kobligk, R. M. Summers, and R. Wiest, "On the interpretability of artificial intelligence in radiology: challenges and opportunities," *Radiology: Artificial Intelligence*, vol. 2, no. 3, p. e190043, 2020. 1

[15] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang, "Concept bottleneck models," in *International Conference on Machine Learning*, 2020, pp. 5338–5348. 1, 2

[16] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626. 1, 2

[17] M. J. Willemink, W. A. Koszek, C. Tan, T. C. Defined, R. L. Defined, M. P. Defined, and B. N. Defined, "Preparing medical imaging data for machine learning," *Radiology*, vol. 295, no. 1, pp. 4–15, 2020. 1

[18] K. Donnelly, "Snomed-ct: The advanced terminology and coding system for ehealth," *Studies in Health Technology and Informatics*, vol. 121, pp. 279–290, 2006. 1

[19] T. Oikarinen, S. Das, L. M. Nguyen, and T.-W. Weng, "Label-free concept bottleneck models," in *International Conference on Learning Representations*, 2023. 1, 2, 3

[20] A. Yan, Y. Wang, Y. Zhong, Z. He, P. Karypis, Z. Wang, C. Dong, A. Gentili, C.-N. Hsu, J. Shang, and J. McAuley, "Robust and interpretable medical image classifiers via concept bottleneck models," *arXiv preprint arXiv:2310.03182*, 2023. 1, 3

[21] S. Jain, A. Agrawal, A. Saporta, S. Q. Truong, D. N. Duong, T. Bui, P. Chambon, Y. Zhang, M. P. Lungren, A. Y. Ng *et al.*, "Radgraph: Extracting clinical entities and relations from radiology reports," in *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*, 2021. 1, 2, 3

[22] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng, "Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific Data*, vol. 6, no. 1, p. 317, 2019. 2, 3

[23] J. C. Denny, "Extracting structured information from free text: challenges and approaches," *AMIA Annual Symposium Proceedings*, vol. 2009, p. 161, 2009. 2

[24] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan, "A simple algorithm for identifying negated findings and diseases in discharge summaries," *Journal of Biomedical Informatics*, vol. 34, no. 5, pp. 301–310, 2001. 2, 3

[25] A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A. Y. Ng, and M. P. Lungren, "Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using bert," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2020, pp. 1500–1519. 2, 3

[26] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, and N. Collier, "Self-alignment pretraining for biomedical entity representations," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, 2021, pp. 4228–4238. 2, 3

[27] O. Bodenreider, "The unified medical language system (umls): integrating biomedical terminology," *Nucleic Acids Research*, vol. 32, no. suppl_1, pp. D267–D270, 2004. 2, 3

[28] Y. Peng, X. Wang, L. Lu, M. Bagheri, R. Summers, and Z. Lu, "Negbio: a high-performance tool for negation and uncertainty detection in radiology reports," *AMIA Summits on Translational Science Proceedings*, vol. 2018, p. 188, 2018. 2, 3

[29] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2097–2106, 2017. 2

[30] P. Rajpurkar, J. Irvin, R. L. Ball, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. P. Langlotz *et al.*, "Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists," *PLoS Medicine*, vol. 15, no. 11, p. e1002686, 2018. 2

[31] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2014. 2

[32] M. Yuksekgonul, M. Wang, and J. Zou, "Post-hoc concept bottleneck models," in *International Conference on Learning Representations*, 2023. 2

[33] M. E. Zarlenga, P. Barbiero, G. Ciravegna *et al.*, "Concept embedding models: Beyond the accuracy-explainability trade-off," in *Advances in Neural Information Processing Systems*, 2022. 2

[34] K. Chauhan, R. Tiwari, J. Freyberg, P. Shenoy, and K. Dvijotham, "Interactive concept bottleneck models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, 2023, pp. 5948–5955. 2

[35] A. Lucieri, M. N. Bajwa, S. A. Braun, M. I. Malik, A. Dengel, and S. Ahmed, "On explainability of deep neural networks for medical image analysis," *arXiv preprint arXiv:2004.08780*, 2020. 2

[36] J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visber *et al.*, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nature Medicine*, vol. 24, no. 9, pp. 1342–1350, 2018. 2

[37] Y. Yang, A. Panagopoulou, S. Sreekumar, I. Chalkidis, M. Yatskar, and C. Callison-Burch, "Language in a bottle: Language model guided concept bottlenecks for interpretable image classification," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19 187–19 197, 2023. 3

[38] D. De Santis, V. Sushko, K. Patel, B. Narayanaswamy, A. Smola, P. Bailis, and T. Kraska, "Visual TCAV: Accurate concept explanations for vision models," in *International Conference on Learning Representations*, 2024. 3

[39] E. Marconato, A. Passerini, and S. Teso, "Glancenets: Interpretable, leak-proof concept-based models," in *Advances in Neural Information Processing Systems*, 2022. 3

[40] B. Kim, K. Gurumoorthy, T. Nguyen, and P. W. Koh, "What changes when concepts shift? robustness analysis of concept bottleneck models," *arXiv preprint arXiv:2402.01234*, 2024. 3

[41] A. Khandelwal and S. Sawant, "Negbert: A transfer learning approach for negation detection and scope resolution," *arXiv preprint arXiv:1911.04211*, 2020. 3

[42] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, 2021, pp. 8748–8763. 3

[43] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, "Contrastive learning of medical visual representations from paired images and text," in *Machine Learning for Healthcare Conference*, 2022, pp. 2–25. 3

[44] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "Medclip: Contrastive learning from unpaired medical images and text," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 3876–3887. 3

[45] S. Zhang, Y. Xu, N. Usuyama, J. Bagher, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, C. Wong *et al.*, "Biomedclip: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs," *arXiv preprint arXiv:2303.00915*, 2023. 3

[46] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2261–2269. 3

[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2015. 3

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778. 3

[49] K. You, J. Gu, J. Ham, B. Park, J. Kim, E. K. Hong, W. Baek, and B. Roh, "Cxr-clip: Toward large scale chest x-ray language-image pre-training," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Springer, 2023, pp. 101–111. 3

[50] A. Rafferty, R. Ramaesh, and A. Rajan, "Xpertxai: An expert-driven concept bottleneck model," *arXiv preprint arXiv:2505.09755*, 2025. 3

[51] D. Alukaev, S. Kiselev, I. Pershin, B. Ibragimov, V. Ivanov, A. Kornaev, and I. Titov, "Cross-modal conceptualization in bottleneck models," *arXiv preprint arXiv:2310.14805*, 2023. 3

[52] V. Sadashivaiah, P. Yan, and J. A. Hendler, "Explaining chest x-ray pathology models using textual concepts," *arXiv preprint arXiv:2407.00557*, 2024. 3

[53] T. F. Chowdhury, V. M. H. Phan, K. Liao, M.-S. To, Y. Xie, A. van den Hengel, J. W. Verjans, and Z. Liao, "Adacbm: An adaptive concept bottleneck model for explainable and accurate diagnosis," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*. Springer, 2024, pp. 35–45. 3

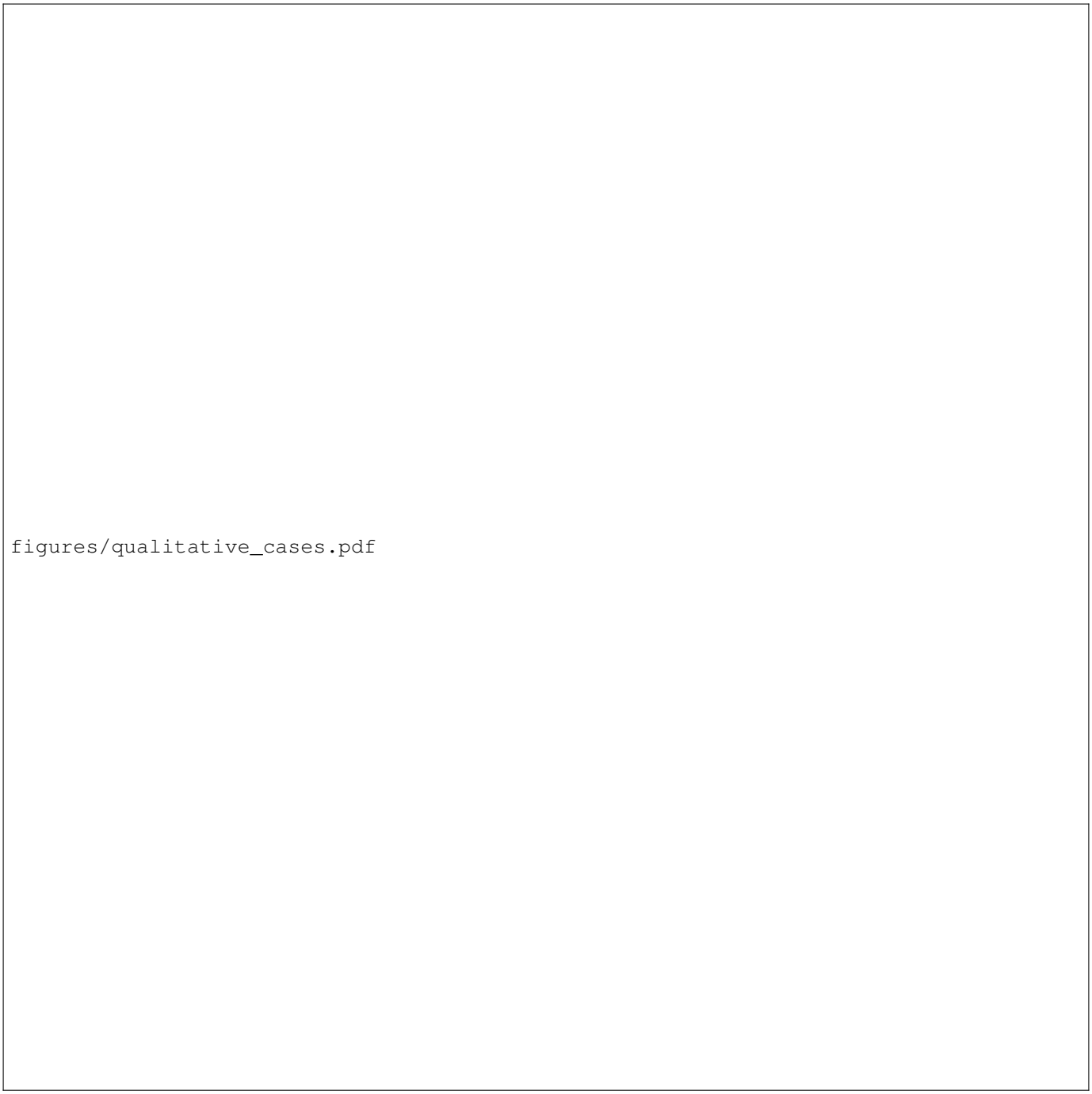

Fig. 7. Qualitative case studies. Each panel shows the input radiograph, region abnormality scores (bar chart), top-5 activated concepts with scores, and ground-truth labels. Case 1: left lower lobe pneumonia. Case 2: cardiomegaly. Case 3: near-normal study. Case 4: pleural effusion with confounding atelectasis. Region-aware explanations align with radiologist reporting conventions.

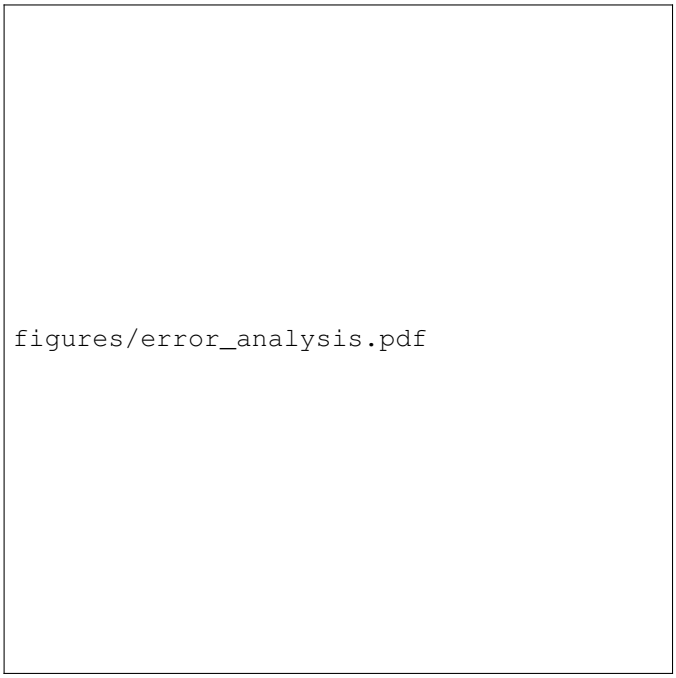

Fig. 8. Error analysis. (a) Region-level confusion matrix showing prediction errors; off-diagonal entries indicate region misclassification rates. (b) False-negative cascade: frequency of missed findings due to incorrect region normality prediction, stratified by finding type.