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

Obadah Habash, Ahmed Alagha, 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 volume alone creates genuine clinical need for systems that reliably flag abnormalities or prioritize urgent cases [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]. But these results have seen limited clinical deployment [10], [11]. Part of this gap reflects concerns about robustness. Models that perform well on internal test sets can fail at external hospitals, sometimes because they exploit institution-specific artifacts rather than disease-related signal [12]. Regulatory, infrastructural, and cultural barriers contribute as well [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 cannot articulate why or point to the cardiothoracic ratio it implicitly computed. It cannot translate that confidence into the criteria clinicians expect, such as measured ratios, anatomical landmarks, or radiographic patterns. Radiologists reason in concepts such as consolidation, air bronchograms, Kerley lines, and costophrenic blunting. A system that cannot engage with this vocabulary offers predictions without a basis for trust or correction [14].

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 become part of the forward pass rather than post-hoc additions through saliency methods [16]. When a CBM predicts pulmonary edema, we can inspect whether it detected cardiomegaly, vascular redistribution, or interstitial opacities and verify that this reasoning aligns with clinical knowledge.

This interpretability 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 prohibitively expensive and does not scale [17], [18]. A single chest radiograph might exhibit dozens of relevant findings across multiple anatomical regions, each requiring expert assessment.

Recent attempts to apply concept-based models to medical imaging have pursued two directions, neither satisfactory for chest radiography. One line of work generates concept vocabularies from large language models, prompting GPT-4 to enumerate radiographic findings and projecting CLIP embeddings onto these concepts [19], [20]. This eliminates manual annotation but introduces new problems. LLM-generated concepts lack grounding in clinical ontologies, may include findings not visually testable from a frontal radiograph, and inherit hallucination tendencies from their source models. A second line integrates clinical knowledge by guiding models to prioritize clinically important concepts through alignment losses [21], [22]. These approaches require expert-provided importance rankings for each concept and have not scaled beyond small concept sets.

We take a different approach and repurpose existing clinical NLP tools as sources of concept supervision. Routine radiology reports already encode rich conceptual supervision. When a radiologist documents "right basilar pneumonia," they have localized disease, described its radiographic pattern, and linked observations to a diagnostic impression. Tools such as RadGraph [23] can extract this structure by parsing reports into entity-relation graphs and are widely used to evaluate report generation quality via RadGraph F1 scores. To our knowledge, they have not been used to supervise concept bottleneck models. Their structured outputs remain confined to evaluation metrics rather than serving as trainable concept targets. 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 [24]. The challenge is transforming this free text into supervision suitable for training concept-based vision models.

This transformation is challenging. Radiology language is dense with abbreviations, implicit negations, and context-dependent qualifications [25]. A finding may be "present," "absent," "unchanged," or "cannot be excluded." These distinctions matter clinically and must be preserved in any derived supervision [26], [27]. Moreover, unmentioned findings are not necessarily absent; radiologists document only what they deem clinically relevant. Linking extracted mentions to standardized terminologies introduces additional complexity, since the same concept may appear in many surface forms and disambiguation requires domain-specific knowledge. Recent advances in clinical natural language processing and biomedical entity linking [28], [23], together with resources such as the Unified Medical Language System (UMLS) [29], now achieve high accuracy on radiology text. But these tools have rarely been combined into pipelines that produce trainable concept banks with assertion status, anatomical context, and ontological grounding.

Prior work has extracted findings from clinical text [30], [7], linked medical entities to ontologies such as UMLS [29], and trained interpretable classifiers on manually curated concept sets [15]. Our approach differs from systems that use reports primarily to derive noisy image-level labels for black-box classifiers, and from CBMs restricted 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. Unlike alignment-loss approaches that require per-concept importance annotations [21], RadCBM encodes clinical knowledge structurally. Ontology grounding via UMLS provides semantic standardization, and hierarchical architecture with multiplicative gating enforces anatomy-first reasoning without additional human input.

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.

Our contributions are as follows:

- We introduce RadCBM, a hierarchical concept bottleneck architecture that organizes concepts by anatomical region and gates region-specific findings through learned region abnormality scores derived from RadGraph-extracted concept locations. Label predictions are linear functions of gated concepts, enforcing clinical consistency (lung findings cannot fire when lungs are predicted normal) without separate region annotations.
- We present a framework that repurposes RadGraph as a source of trainable concept supervision rather than solely an evaluation metric. 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.

II. Related Work

A. Deep Learning for Chest Radiography

Progress in chest radiograph interpretation has been driven by large-scale datasets including ChestX-ray14 [31], CheXpert [7], and MIMIC-CXR [24]. Classification architectures have evolved from DenseNet-based models [6] to Vision Transformers [9], with recent foundation models such as Ark+ demonstrating strong generalization to rare and novel diseases [32]. Parallel work has pursued vision-language pre-training: ConVIRT [33], MedCLIP [34], and BiomedCLIP [35] align radiograph and report embeddings for zero-shot transfer. These approaches achieve high accuracy but produce entangled representations that lack explicit clinical semantics, motivating concept-based alternatives.

B. Concept Bottleneck Models

Koh et al. [15] introduced Concept Bottleneck Models (CBMs), which route predictions through interpretable intermediate concepts. Subsequent work has relaxed the strict bottleneck via concept embeddings [36], enabled post-hoc retrofitting of pretrained networks [37], and supported test-time concept intervention [38]. Coarse-to-fine architectures tie global predictions to localized findings [39]. The persistent bottleneck is supervision: label-free CBMs [18], [40] align CLIP representations with concept predictors to avoid manual annotation, while visual concept filtering [20] prunes LLM-generated vocabularies to visually grounded subsets. However, Debole et al. [41] show that VLM-derived supervision diverges substantially from expert annotations and that concept accuracy does not correlate with downstream task performance, underscoring the need for higher-quality concept sources. In radiology, AHIVE [42] organizes visual features by anatomical region for report retrieval. Our work addresses concept quality directly by constructing the concept bank from structured report extraction and ontology grounding rather than VLM weak supervision.

C. Concept Extraction from Radiology Reports

Rule-based extractors such as NegBio [30] and the CheXpert labeler [7] identify findings and negation via pattern matching; CheXbert [27] improved accuracy with BERT fine-tuning. RadGraph [23] extended extraction to entity-relation graphs over observations and anatomy. RadGraph-XL [43] scaled annotation to 2,300 reports across four modalities; RadGraph2 [44] added temporal change tracking. For entity linking, SapBERT [28] achieves strong performance grounding mentions to UMLS [29]. These tools are widely used for evaluating generated reports via RadGraph F1, but their structured outputs have not been repurposed as CBM supervision. We bridge this gap.

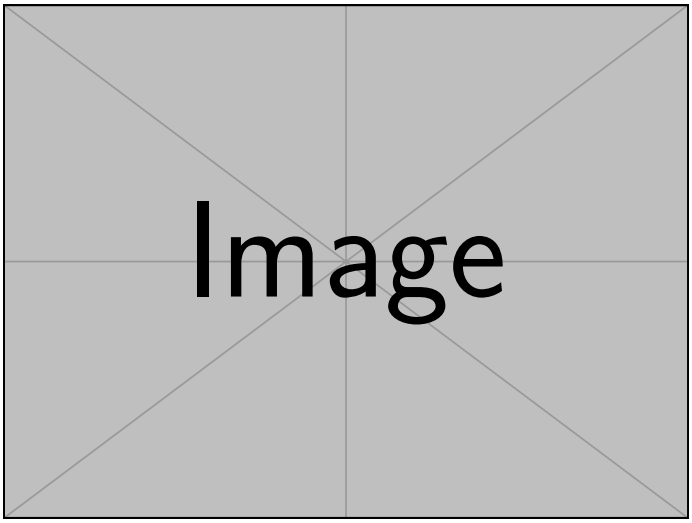

Fig. 1. RadCBM architecture overview. Stage 1 trains a concept predictor with anatomical region gating. Stage 2 trains a linear diagnosis head on gated concepts.

III. Method

RadCBM predicts diagnoses through a two-stage concept bottleneck with anatomical gating (Fig. 1). Stage 1 trains a concept predictor (with a frozen pretrained image encoder) to predict both fine-grained concepts and coarse region abnormality scores from report-derived supervision. Stage 2 trains a linear diagnosis head on region-gated concepts, with the concept predictor frozen. We use i for studies, k for concepts, r for regions, and j for diagnosis labels. We first describe concept bank construction, then detail the model architecture and training procedure.

A. Concept Bank Construction

We convert free-text radiology reports into structured concept supervision through entity extraction, ontology linking, and vocabulary construction.

Entity extraction. RadGraph-XL [23] parses report findings and impression sections into entity mentions, each labeled with an assertion status: *present*, *absent*, or *uncertain*. Modifier spans linked by RadGraph relations are retained, as they often encode laterality or coarse location (e.g., "left lower lobe," "bilateral").

Ontology linking. Extracted mentions vary in surface form: "opacity," "opacification," and "opacities" may refer to the same finding. We standardize terminology by linking each mention to a UMLS Concept Unique Identifier (CUI). Specifically, we embed mentions using SapBERT [28] and retrieve the nearest neighbor from an index of SNOMED-CT synonyms [45]. To reduce off-domain matches, we restrict candidates to clinically relevant semantic types: observations (T047: Disease/Syndrome, T046: Pathologic Function, T033: Finding) and anatomy (T017: Anatomical Structure, T023: Body Part, T029: Body Location). Matches with cosine similarity below 0.8 are discarded.

Vocabulary construction. Linked concepts are aggregated across all studies. We retain concepts exceeding a frequency threshold with at least one positive assertion, and filter uninformative normality phrases (e.g., "unremarkable," "no acute findings") by name matching. The resulting concept bank $\mathcal{C} = \{c_k\}_{k=1}^{K}$ contains 1,312 ontology-grounded findings, two orders of magnitude larger than the 14-class vocabularies in standard benchmarks.

Mention masking and uncertainty. Because radiologists document selectively, missing mentions cannot be treated as negative labels. For each study i, we construct a mention mask vector $m_i \in \{0,1\}^K$ indicating which concepts were explicitly asserted. Unmentioned concepts ($(m_i)_k$=0) are excluded from the loss. For mentioned concepts, we derive targets $t_i \in \{0, 0.5, 1\}^K$: absent assertions map to 0, present to 1, and uncertain to 0.5 as a soft target.

Anatomical grouping. Each concept is assigned to one of six anatomical regions $\mathcal{R}$: lung, pleura, heart, mediastinum, bone, or other. Assignment uses location strings extracted from report modifiers and name heuristics (e.g., "pleur-" $\rightarrow$ pleura, "pulmon-" $\rightarrow$ lung). This defines a fixed parent mapping $g : \{1, \ldots, K\} \rightarrow \mathcal{R}$ from concept index k to its region. Region-level targets are obtained by max-pooling over constituent concepts:

$$(\tilde{z}_i)_r = \max_{k:\, g(k)=r} (t_i)_k. \tag{1}$$

We similarly define a region mask $\tilde{m}_i \in \{0,1\}^{|\mathcal{R}|}$ with entries $(\tilde{m}_i)_r = \max_{k:\, g(k)=r}(m_i)_k$; regions containing no mentioned concepts are masked from region supervision.

B. Model Architecture

Stage 1: Image to concepts and regions. Given an image x_i, a frozen pretrained vision encoder extracts features $h_i = f_\theta(x_i)$. A two-layer MLP produces concept logits $s_i \in \mathbb{R}^K$ with probabilities $\hat{c}_i = \sigma(s_i)$.

Region gates are derived from concept probabilities rather than image features, so that gating reflects the model's own concept-level evidence. Region logits $u_i = W_r \hat{c}_i$ are converted to gate probabilities:

$$\hat{z}_i = \epsilon + (1 - \epsilon)\, \sigma(u_i/\tau), \tag{2}$$

where τ is a temperature parameter and ϵ is a floor preventing full gate closure. Each concept is then gated by its parent region:

$$(\hat{c}_i^{\text{gated}})_k = (\hat{z}_i)_{g(k)} \cdot (\hat{c}_i)_k. \tag{3}$$

This enforces anatomical consistency: a pleural finding cannot contribute to predictions when the pleura gate is low.

The Stage 1 objective combines mention-masked concept and region losses:

$$\mathcal{L}_{\text{stage1}} = \mathcal{L}_{\text{concept}} + \lambda_r \mathcal{L}_{\text{region}}, \tag{4}$$

where

$$\mathcal{L}_{\text{concept}} \coloneqq \frac{1}{\sum_{i=1}^{N} \|m_i\|_1} \sum_{i=1}^{N} \|m_i \odot \text{BCE}(s_i, t_i)\|_1,$$

$$\mathcal{L}_{\text{region}} \coloneqq \frac{1}{\sum_{i=1}^{N} \|\tilde{m}_i\|_1} \sum_{i=1}^{N} \|\tilde{m}_i \odot \text{BCE}(u_i, \tilde{z}_i)\|_1,$$

where N is the number of studies, $\text{BCE}(\cdot, \cdot)$ denotes binary cross-entropy on logits (applied element-wise), $\odot$ denotes element-wise multiplication, and $\|\cdot\|_1$ sums vector entries. If a denominator is zero, the corresponding term is omitted.

Stage 2: Gated concepts to diagnoses. With the vision encoder and concept head frozen, we train a diagnosis head on the predicted (gated) concept probabilities. To preserve interpretability, we use a bias-free linear layer:

$$\ell_{ij} = \sum_k W_{jk} \cdot (\hat{c}_i^{\text{gated}})_k. \tag{5}$$

The contribution of concept k to diagnosis j is directly readable as $W_{jk} \cdot (\hat{c}_i^{\text{gated}})_k$: positive weights indicate supportive findings, negative weights indicate contradictory ones. Omitting the bias ensures the model cannot predict disease without activated concepts.

We train with binary cross-entropy on CheXpert-style multi-labels $y_i \in \{0, 1, -1\}^L$, where L is the number of diagnosis labels and -1 denotes uncertainty. By default, uncertain labels are treated as missing and excluded from the loss; we report sensitivity analyses with alternative uncertainty mappings in the supplement.

C. Training and Inference

Training. Both stages are trained with Adam using learning rate search and early stopping on validation performance. All experiments use predefined train/validation/test splits; images, concept labels, and disease labels are aligned by study identifier to prevent leakage. Hyperparameters (τ, ϵ, λ_r) are selected via validation; details and sensitivity analyses are provided in the supplement.

Inference. Given a test image x, we compute concept probabilities $\hat{c}$ and region gates $\hat{z}$, apply gating via Eq. (3), and obtain diagnosis probabilities $\hat{y} = \sigma(\ell)$ from Eq. (5). For interpretability, we report the top-k concept contributions $W_{jk} \cdot (\hat{c}^{\text{gated}})_k$ per diagnosis and threshold activations at δ=0.5 when summarizing.

Algorithm 1 summarizes the full procedure.

Algorithm 1 RadCBM Training and Inference

Require: Image x, concept targets $t \in [0, 1]^K$, mention mask $m \in \{0, 1\}^K$, region map g, disease labels y
Ensure: Diagnosis probabilities $\hat{y}$, concept contributions

Stage 1: Learn concept predictor

$h \leftarrow f_\theta(x)$ ▷ *Frozen vision encoder*
$\hat{c} \leftarrow \sigma(\text{MLP}_\phi(h))$ ▷ *Concept probabilities*
$\hat{z} \leftarrow \epsilon + (1 - \epsilon)\, \sigma(W_r \hat{c} \,/\, \tau)$ ▷ *Region gates*
$(\hat{c}^{\text{gated}})_k \leftarrow (\hat{z})_{g(k)} \cdot (\hat{c})_k \quad \forall k$ ▷ *Gated concepts*
Minimize $\mathcal{L}_{\text{stage1}}$ over ϕ, W_r ▷ *Mention-masked*

Stage 2: Learn diagnosis head

Freeze ϕ, W_r
$\ell_j \leftarrow \sum_k W_{jk} \cdot (\hat{c}^{\text{gated}})_k \quad \forall j$ ▷ *Linear, no bias*
Minimize $\mathcal{L}_{\text{label}}$ over W ▷ *Masked BCE on y*

Inference

Compute $\hat{c}, \hat{z}, \hat{c}^{\text{gated}}$ as above
$\hat{y} \leftarrow \sigma(\ell)$
return $\hat{y}$, contributions $\{W_{jk} \cdot (\hat{c}^{\text{gated}})_k\}_{j,k}$

IV. Results

A. Experimental Setup

1) Datasets: We evaluate on five chest radiograph benchmarks spanning in-domain and external validation. **MIMIC-CXR** [24] contains 377,110 radiographs from 65,379 patients with associated radiology reports; we use the official train/validation/test splits stratified by patient. **CheXpert Plus** builds on CheXpert [7]; we evaluate on the radiologist-labeled expert subset. **VinDr-CXR** [46] provides radiologist annotations for 28 findings, **RSNA Pneumonia** [47] provides pneumonia detection labels with bounding boxes, and **NIH ChestX-ray14** [31] provides 14 disease labels mined from reports.

CheXpert Plus (expert subset), VinDr-CXR, and RSNA Pneumonia provide radiologist-annotated evaluation labels, while NIH ChestX-ray14 labels are report-derived. We emphasize performance on radiologist-labeled subsets as primary evidence of clinical correctness and treat report-derived targets as complementary large-scale evidence.

2) Concept Bank Construction: We extract concepts exclusively from MIMIC-CXR training reports using RadGraph [23], yielding 23,452 unique observation-anatomy pairs. After UMLS normalization, semantic type filtering, and frequency thresholding (minimum 50 occurrences), the final vocabulary contains 1,312 region-specific concepts organized into six anatomical regions: lung (259 concepts), heart (143 concepts), pleura (69 concepts), mediastinum (116 concepts), bone (144 concepts), and other (581 concepts). Assertion status (present, absent, uncertain) is preserved for each concept mention.

3) Implementation Details: We implement RadCBM with a frozen radiology-pretrained MedCLIP vision backbone [34] (Swin-T). Unless stated otherwise, all CBM comparisons in this section use the same MedCLIP transformer backbone (Swin-T) frozen to isolate bottleneck design effects. We additionally report black-box VLM baselines using CXR-CLIP [48] and

CheXzero [49]. Images are resized to the backbone's native resolution and normalized accordingly. We apply standard augmentations during training: random horizontal flipping, rotation ($\pm 10°$), and color jittering.

Models are trained using Adam [50] with learning rate 10^{-4}, batch size 32, and early stopping based on validation macro AUC (patience 10 epochs). We set region loss weight $\lambda_r = 1.0$, temperature $\tau = 1.0$, and gate floor $\epsilon = 0.0$ (no clamping); sensitivity analyses are in the supplement. All experiments were conducted on an NVIDIA GeForce RTX 3080 GPU (16GB). We report results averaged over 3 random seeds.

4) Baselines and Comparison Protocol: We compare against concept bottleneck models (CBMs) and black-box baselines.

CBM baselines. (1) **Post-hoc CBM** [37], which retrofits concept bottlenecks onto pretrained models; **LaBo** [51], which constructs text-defined bottlenecks with linear concept-to-class predictors (Following LaBo's prompts and 500-sentences/class budget, we use OpenAI `gpt-3.5-turbo` to generate candidate sentences and extract short concepts; this substitutes for LaBo's deprecated GPT-3 generator and unreleased fine-tuned T5 extractor, and we keep LaBo's cleaning heuristics unchanged); (3) **AdaCBM** [52], which adds an adaptive module to reduce domain mismatch; and (4) **C2F-CBM** [39], which builds two-level bottlenecks with coarse-to-fine prediction.

Black-box baselines. Supervised CNNs (ResNet-50, DenseNet-121) [53] and vision-language models used as black-box encoders (MedCLIP, CXR-CLIP, CheXzero).

Comparison protocol. To ensure fair comparison, we follow recent recommendations for evaluating VLM-CBMs [54]:

- All CBMs within a comparison use the same frozen vision backbone.
- For concept-level evaluation (Table II), all methods are evaluated on the same 1,312-concept target set. For LaBo, we use an ontology-aligned variant ("LaBo (fixed vocab)") that takes our concept bank as input.
- Hyperparameters are tuned per method via validation-based early stopping with a fixed search budget.
- Train/validation/test splits and uncertainty handling are identical across methods.

All supervised baselines and CBMs are trained on MIMIC-CXR train/val and evaluated on MIMIC-CXR test and external datasets without retraining; for RSNA (binary pneumonia), we fit a dataset-specific pneumonia head when the method lacks that label. Zero-shot VLMs are evaluated directly on each test set. For label-level evaluation (Tables I, III), each CBM method uses its native concept source, as the concept bank is part of the method's contribution. For interpretability metrics (Table IV), we evaluate only intrinsic CBMs where the bottleneck mediates predictions; post-hoc CBMs are excluded since concept interventions do not affect their underlying predictor.

5) Evaluation Metrics: **Classification performance** is reported using per-label and macro-averaged AUC-ROC on the five CheXpert competition labels (Atelectasis, Cardiomegaly, Consolidation, Edema, Pleural Effusion). Full 14-label results are in the supplement. When thresholded metrics are reported, per-label thresholds are tuned on MIMIC-CXR validation and fixed for all test sets.

Concept quality is assessed on the shared 1,312-concept bank using macro AUC-ROC and macro AUPRC, reported overall and on rare concepts (50–200 training occurrences). We compute concept metrics only on explicitly asserted mentions; unmentioned concepts are unlabeled. Macro averages include only concepts with at least one labeled positive and one labeled negative example.

Interpretability is evaluated via three metrics: (1) *Intervention faithfulness*: Pearson correlation between predicted concept contribution ($W_{jk} \cdot \hat{c}_k^{\text{gated}}$) and observed label change upon setting $\hat{c}_k$ to 0 or 1. For the linear head, this correlation is 1.0 by construction. (2) *Plausibility*: fraction of activated findings ($\hat{c}_k > 0.5$) whose parent region gate exceeds 0.5. (3) *Implausible activation rate*: fraction of finding activations occurring when the parent region gate is below 0.3. (4) *Region consistency*: agreement between region gates and max-pooled concept activations (Eq. 6 in supplement).

B. Classification Performance

Table I presents classification performance on the five CheXpert competition labels. RadCBM is competitive with CBM baselines while providing interpretable concept-mediated predictions. On MIMIC-CXR, RadCBM achieves a macro AUC of 0.794 ± 0.001, outperforming intrinsic CBMs (LaBo, AdaCBM) and approaching the post-hoc CBM, while remaining competitive with the supervised DenseNet-121 baseline (0.684). Hierarchical gating yields a small macro AUC gain over the flat variant on MIMIC-CXR (0.794 vs. 0.788), with negligible per-label differences ($\leq$0.001) for Pleural Effusion and Edema.

Among CBM approaches, methods relying on small or automatically generated concept vocabularies 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 on the shared 1,312-concept target set. We report concept quality only for intrinsic CBMs; post-hoc CBMs are omitted since their concept scores are not trained to match a target concept bank. RadCBM achieves the highest overall concept AUC (0.693) and AUPRC (0.861). Rare-concept AUC is lower than the flat variant (0.507 vs. 0.641), suggesting gating trades rare sensitivity for stronger overall calibration.

D. External Validation

Table III reports generalization to external benchmarks. For multi-label datasets, we report macro AUC over the five CheXpert competition labels using dataset-specific mappings; for RSNA we report binary pneumonia AUC. Per-label thresholds tuned on MIMIC-CXR validation are applied without further tuning. For RSNA, we fit a binary pneumonia head when the method does not include a native pneumonia label.

RadCBM generalizes competitively across all benchmarks, with smaller performance drops than black-box baselines on distribution shift (VinDr-CXR, NIH). This suggests that ontology-grounded concepts provide more transferable intermediate representations than end-to-end learned features.

TABLE I
CLASSIFICATION PERFORMANCE (AUC-ROC) ON MIMIC-CXR TEST SET AND CHEXPERT PLUS EXPERT SUBSET FOR FIVE COMPETITION LABELS. BEST IN **BOLD**, SECOND-BEST UNDERLINED. ALL CONCEPT-BASED METHODS SHARE THE SAME FROZEN BACKBONE (MEDCLIP SWIN-T). RESULTS ARE MEAN ± STD OVER 3 SEEDS FOR SEEDED METHODS; SINGLE-RUN BASELINES ARE SHOWN WITHOUT STD.

Method	Type	Atelect.	Cardiom.	Consolid.	Edema	Pl. Eff.	Macro
MIMIC-CXR Test Set							
ResNet-50	CNN	0.6623	0.7145	0.6674	0.7552	0.8038	0.7207
DenseNet-121	CNN	0.6243	0.7014	0.5898	0.7540	0.7497	0.6839
MedCLIP (Swin-T)	VLM	0.7387	0.7481	0.7662	0.8457	0.8876	0.7973
CXR-CLIP (Swin-T)	VLM	0.4494	0.5843	0.5156	0.6409	0.5058	0.5392
CheXzero (ViT-B/32)	VLM	0.6531	0.7249	0.7023	0.8245	0.8423	0.7494
Post-hoc CBM	CBM	0.7662 ± 0.0000	0.7746 ± 0.0000	0.7203 ± 0.0001	0.8577 ± 0.0000	0.8951 ± 0.0000	0.8028 ± 0.0000
LaBo	CBM	0.6345 ± 0.0003	0.7363 ± 0.0000	0.5329 ± 0.0001	0.7194 ± 0.0002	0.8844 ± 0.0000	0.7015 ± 0.0001
AdaCBM	CBM	0.7410 ± 0.0048	0.7771 ± 0.0010	0.6378 ± 0.0018	0.8267 ± 0.0052	0.8906 ± 0.0006	0.7747 ± 0.0018
C2F-CBM	H-CBM	0.7684 ± 0.0049	0.7441 ± 0.0066	0.7386 ± 0.0088	0.8574 ± 0.0021	0.8757 ± 0.0020	0.7968 ± 0.0048
RadCBM (flat)	CBM	0.7588	0.7460	0.7201	0.8464	0.8711	0.7885 ± 0.0015
RadCBM	H-CBM	0.7622	0.7495	0.7380	0.8471	0.8716	0.7937 ± 0.0009
CheXpert Plus Expert Subset							
ResNet-50	CNN	0.5581	0.4777	0.4197	0.5224	0.5511	0.5058
DenseNet-121	CNN	0.5288	0.4215	0.4921	0.5957	0.5664	0.5209
MedCLIP (Swin-T)	VLM	0.4938	0.5921	0.6271	0.4924	0.4875	0.5386
CXR-CLIP (Swin-T)	VLM	0.5048	0.4351	0.5977	0.5048	0.5007	0.5086
CheXzero (ViT-B/32)	VLM	0.5396	0.4912	0.5841	0.5535	0.5483	0.5433
Post-hoc CBM	CBM	0.5781 ± 0.0002	0.3779 ± 0.0005	0.4170 ± 0.0004	0.5330 ± 0.0002	0.5299 ± 0.0001	0.4872 ± 0.0000
LaBo	CBM	0.4452 ± 0.0004	0.4812 ± 0.0001	0.4654 ± 0.0000	0.5342 ± 0.0001	0.4880 ± 0.0000	0.4828 ± 0.0001
AdaCBM	CBM	0.5589 ± 0.0029	0.3598 ± 0.0012	0.5856 ± 0.0091	0.5452 ± 0.0019	0.5388 ± 0.0017	0.5176 ± 0.0027
C2F-CBM	H-CBM	0.5889 ± 0.0020	0.4303 ± 0.0037	0.4772 ± 0.0084	0.5897 ± 0.0010	0.5502 ± 0.0023	0.5273 ± 0.0021
RadCBM (flat)	CBM	0.5842	0.4249	0.5001	0.5865	0.5407	0.5273 ± 0.0057
RadCBM	H-CBM	0.5836	0.4232	0.5006	0.5844	0.5406	0.5265 ± 0.0011

TABLE II
CONCEPT PREDICTION QUALITY ON MIMIC-CXR TEST SET (SHARED 1,312-CONCEPT BANK). †LABO EVALUATED WITH ONTOLOGY-ALIGNED VARIANT. RESULTS: MEAN ± STD OVER 3 SEEDS WHEN AVAILABLE.

Method	AUC	AUPRC	Rare AUC	Rare AUPRC
LaBo†	0.474 ± 0.000	0.778 ± 0.000	0.496 ± 0.000	0.728 ± 0.000
AdaCBM	0.622 ± 0.003	0.854 ± 0.001	0.614 ± 0.011	0.781 ± 0.002
C2F-CBM	0.542 ± 0.001	0.810 ± 0.001	0.446 ± 0.018	0.668 ± 0.005
RadCBM (flat)	0.517	0.804	0.641	0.782
RadCBM	0.693	0.861	0.507	0.730

TABLE III
EXTERNAL VALIDATION (AUC-ROC). MULTI-LABEL COLUMNS: MACRO AUC OVER FIVE CHEXPERT LABELS; RSNA: BINARY PNEUMONIA AUC. POST-HOC CBM IS OMITTED SINCE ITS PREDICTIONS MATCH ITS UNDERLYING BLACK-BOX MODEL BY CONSTRUCTION. RSNA VALUES FOR C2F ARE OMITTED (—) DUE TO MISSING PNEUMONIA HEADS.

Method	Type	MIMIC	CheXpert Plus	VinDr-CXR	NIH	RSNA
ResNet-50	CNN	0.721	0.506	0.867	0.817	0.892
DenseNet-121	CNN	0.684	0.521	0.854	0.647	0.772
MedCLIP	VLM	0.797	0.539	0.910	0.789	0.831
CXR-CLIP (Swin-T)	VLM	0.539	0.509	0.841	0.744	0.779
CheXzero (ViT-B/32)	VLM	0.749	0.543	0.870	0.765	0.839
LaBo	CBM	0.702	0.483	0.927	0.763	0.787
AdaCBM	CBM	0.775	0.518	0.935	0.762	0.898
C2F-CBM	H-CBM	0.797	0.527	0.926	0.765	—
RadCBM (flat)	CBM	0.788	0.527	0.927	0.757	0.873
RadCBM	H-CBM	0.794	0.526	0.927	0.761	0.873

E. Interpretability

Table IV evaluates whether concept-based explanations support faithful interventions and clinically plausible activations.

TABLE IV
INTERPRETABILITY METRICS ON MIMIC-CXR (INTRINSIC CBMS ONLY). †LABO EVALUATED WITH ONTOLOGY-ALIGNED VARIANT. RESULTS: MEAN ± STD OVER 3 SEEDS.

Method	Interv. Faith. ↑	Plaus. ↑	Implaus. Rate ↓	Region Cons. ↑
LaBo†	0.752 ± 0.033	1.000 ± 0.000	0.000 ± 0.000	0.511 ± 0.005
AdaCBM	0.864 ± 0.015	1.000 ± 0.000	0.000 ± 0.000	0.531 ± 0.016
C2F-CBM	0.997 ± 0.001	0.981 ± 0.002	0.017 ± 0.002	—
RadCBM (flat)	0.999 ± 0.000	0.987 ± 0.000	0.001 ± 0.000	—
RadCBM	1.000 ± 0.000	1.000 ± 0.000	0.000 ± 0.000	0.847 ± 0.026

We report metrics only for intrinsic CBMs where the bottleneck mediates label predictions.

RadCBM achieves perfect intervention faithfulness by construction (the linear head ensures predicted and observed effects match exactly). Hierarchical gating reduces the implausible activation rate from 0.11% (flat) to 0.00%, indicating that region gates successfully suppress findings in anatomically inactive regions. Region consistency (0.847) confirms that gates align with pooled concept evidence.

F. Ablation Study

We report ablations isolating key design choices in the supplement (Table VI). Briefly, mention masking and assertion-aware targets improve concept quality by reducing label noise, hierarchical gating improves plausibility with minimal impact

on classification AUC, and conservative soft-gating ($\epsilon > 0$) mitigates cascading failures where missed region predictions suppress all constituent findings.

V. Conclusion

RadCBM turns routine radiology reports into large-scale concept supervision and aligns model reasoning with how radiologists read chest X-rays...

Appendix A
Supplementary Material

A. Evaluation Label Provenance

Table V summarizes which benchmarks provide radiologist-annotated evaluation labels versus report-derived labels.

TABLE V
EVALUATION LABEL SOURCE AT TEST TIME.

Benchmark	Label Source	Level	Label Set
MIMIC-CXR	radiologist-annotated (reports; test split)	study	CheXpert-14
CheXpert Plus	radiologist-annotated	study	CheXpert-14
VinDr-CXR	radiologist-annotated	image+bbox	VinDr-28 (mapped to CheXpert-5)
RSNA Pneumonia	radiologist-annotated	image+bbox	Pneumonia
NIH ChestX-ray14	report-derived	image	NIH-14

For MIMIC-CXR, we evaluate using radiologist-annotated report labels released with the PhysioNet MIMIC-CXR-JPG test split (CheXpert-14 categories). These are report annotations rather than independent radiologist re-reads of the images. NIH ChestX-ray14 labels are report-derived; we treat them as complementary evidence and emphasize radiologist-labeled subsets as primary validation.

B. Region Consistency Metric

We quantify alignment between coarse region gates and fine concept evidence using:

$$\mathrm{RC} = 1 - \frac{1}{N|\mathcal{R}|} \sum_{i=1}^{N} \sum_{r \in \mathcal{R}} \left| (\hat{z}_i)_r - \max_{k:g(k)=r} (\hat{c}_i)_k \right|. \quad (6)$$

Values near 1 indicate that region gates faithfully summarize the underlying concept activations.

C. Ablation Study

Table VI presents an incremental ablation isolating key design choices. We add components to a base RadCBM model while keeping the evaluation protocol fixed (thresholds tuned on MIMIC-CXR validation, then frozen).

TABLE VI
ABLATION STUDY ON MIMIC-CXR TEST SET. RESULTS AVERAGED OVER 3 SEEDS.

Configuration	Macro AUC	Concept AUC	Plaus.	Interv. Faith.
RadCBM (base)	**.XXX**	**.XXX**	**.XXX**	**.XX**
+ Mention masking	**.XXX**	**.XXX**	**.XXX**	**.XX**
+ Assertion-aware targets	**.XXX**	**.XXX**	**.XXX**	**.XX**
+ Hierarchical gating	**.XXX**	**.XXX**	**.XXX**	**.XX**
+ Conservative soft-gating ($\epsilon > 0$)	**.XXX**	**.XXX**	**.XXX**	**.XX**
RadCBM (full)	**.XXX**	**.XXX**	**.XXX**	**.XX**

Mention masking excludes unmentioned concepts from supervision rather than treating them as negatives, improving concept AUC by reducing label noise. **Assertion-aware targets** maps present/absent/uncertain assertions to 1/0/0.5 rather than binary labels, providing softer supervision for ambiguous findings. **Hierarchical gating** introduces region-level gates that suppress anatomically implausible concept activations. **Conservative soft-gating** sets $\epsilon > 0$ to prevent gates from fully closing, avoiding cascading failures where a missed region prediction suppresses all constituent findings.

D. Region-Level Performance

Table VII reports performance decomposed by anatomical region. Region AUC measures binary abnormality detection using surrogate targets obtained by max-pooling concept assertions per region. Finding AUC measures concept prediction within each region.

TABLE VII
REGION-LEVEL PERFORMANCE ON MIMIC-CXR TEST SET. RESULTS AVERAGED OVER 3 SEEDS. THE "OTHER" CATEGORY (581 CONCEPTS) IS EXCLUDED AS IT AGGREGATES HETEROGENEOUS FINDINGS WITHOUT CLEAR ANATOMICAL LOCALIZATION.

Region	#Concepts	Region AUC	Finding AUC	Prevalence (%)
Lung	259	0.684±0.000	0.691±0.000	92.1
Heart	143	0.822±0.000	0.770±0.000	81.1
Pleura	69	0.727±0.000	0.698±0.000	81.9
Mediastinum	116	0.592±0.000	0.701±0.000	89.7
Bone	144	0.702±0.000	0.676±0.000	78.6
Overall	1,312	0.678±0.000	0.693±0.000	—

E. Learned Concept-Label Relationships

Missing figure:
`figures/concept_label_weights.pdf`

Fig. 2. Learned concept-to-label weights from the linear diagnosis head. Each row shows the top-5 positive and top-5 negative concept contributions for one CheXpert label.

F. Concept Bank Statistics

Fig. 3. Concept bank statistics. (a) Concept frequency distribution (log scale); vertical lines indicate CheXpert-14 concept positions. (b) Hierarchical organization by anatomical region. (c) Vocabulary coverage comparison with prior work.

G. Effect of Hierarchical Gating

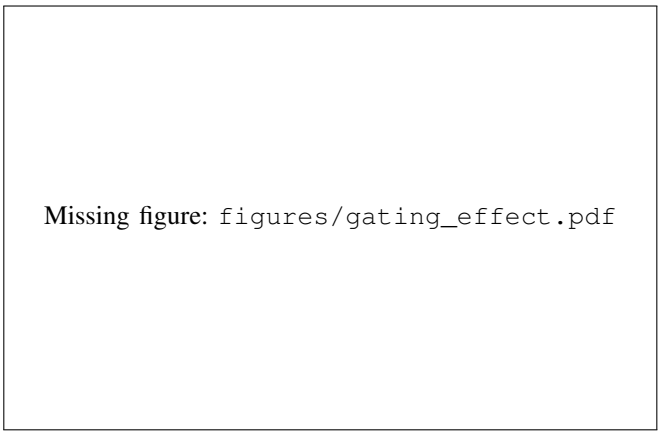

Fig. 4. Effect of hierarchical gating. (a) Region abnormality score versus mean finding activation for flat vs hierarchical variants. (b) Distribution of finding activations stratified by region gate status.

H. Intervention Faithfulness Analysis

Fig. 5. Intervention faithfulness. (a) Label probability as a function of concept activation. (b) Predicted versus observed label change upon concept intervention.

I. Hyperparameter Sensitivity

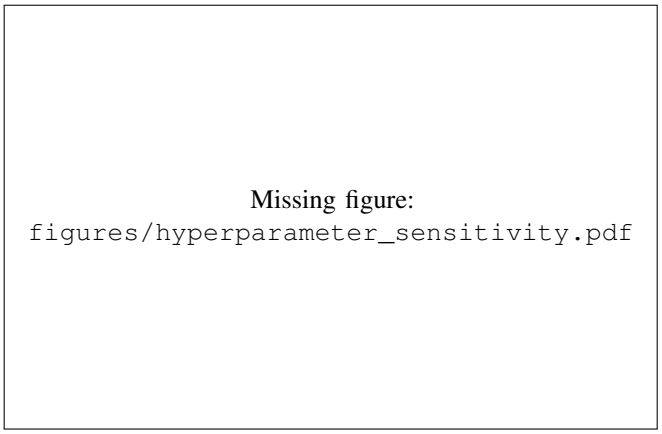

Fig. 6. Sensitivity to region loss weight λ_r. Validation macro AUC remains stable across $\lambda_r \in [0.01, 1.0]$.

J. Concept AUC by Frequency

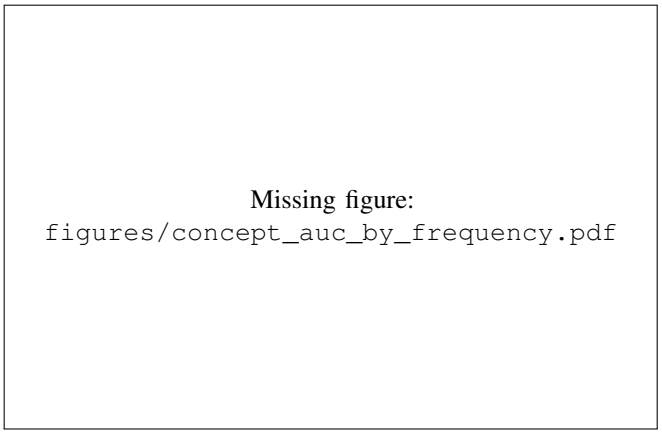

Fig. 7. Concept AUC stratified by training set frequency. Hierarchical gating provides larger gains for rare concepts.

K. Qualitative Case Studies

L. Cross-Dataset Generalization

TABLE VIII
CROSS-DATASET GENERALIZATION. MODELS TRAINED ON ONE DATASET AND EVALUATED ON ANOTHER. Δ INDICATES CHANGE FROM IN-DOMAIN PERFORMANCE.

Method	**Train → Test**	**Macro AUC**	Δ
DenseNet-121	MIMIC → CheXpert	0.521	-23.8%
RadCBM	MIMIC → CheXpert	0.526	-33.7%
DenseNet-121	CheXpert → MIMIC	0.684	$+31.3\%$
RadCBM	CheXpert → MIMIC	0.659	$+18.9\%$

M. Computational Requirements

TABLE IX
COMPUTATIONAL REQUIREMENTS ON MIMIC-CXR. PARAMS COUNT TRAINABLE PARAMETERS ONLY (FROZEN MEDCLIP BACKBONES ARE SHARED ACROSS CBMS AND EXCLUDED). INFERENCE MEASURED ON NVIDIA RTX 3080 WITH BATCH SIZE 1 USING THE FULL PIPELINE.

Method	Params (M)	Inference (ms)	Training (GPU-hrs)
DenseNet-121	7.0	9.9	0.0
AdaCBM	0.5	9.5	0.02
RadCBM (flat)	4.4	12.7	2.1
RadCBM	2.7	16.6	2.3

N. Error Analysis

Missing figure: `figures/error_analysis.pdf`

Fig. 9. Error analysis. (a) Region-level confusion matrix. (b) False-negative cascade: missed findings due to incorrect region normality prediction.

O. Calibration

TABLE X
CALIBRATION ON CHEXPERT PLUS EXPERT SUBSET. LOWER IS BETTER.

Method	ECE ↓	Brier ↓
DenseNet-121	0.287	0.291
MedCLIP	0.174	0.237
RadCBM	0.430	0.402

Missing figure: `figures/reliability_diagram.pdf`

Fig. 10. Reliability diagram on CheXpert Plus expert subset.

P. Rare-Label Performance

TABLE XI
PR-AUC ON RARE LABELS (CHEXPERT PLUS EXPERT SUBSET). LUNG LESION AND PLEURAL OTHER HAVE NO POSITIVES IN THIS SUBSET, SO VALUES ARE REPORTED AS — AND EXCLUDED FROM THE MACRO AVERAGE.

Method	Fracture	Pneumothorax	Pneumonia	Lung Lesion	Pleural Other	Macro
DenseNet-121	0.69	0.26	0.92	—	—	0.623
MedCLIP	0.54	0.27	1.00	—	—	0.602
RadCBM	0.79	0.25	0.92	—	—	0.650

Q. Uncertainty Handling Protocol

Unless otherwise stated, we map labeler outputs to {positive, negative, uncertain}. For training labels derived from reports, we optionally use an ensemble of labelers (CheXpert, CheXbert, NegBio); disagreements are marked uncertain. For disease labels, uncertain values are treated as missing and excluded from loss and evaluation. For concept targets, uncertain assertions receive soft targets (0.5). Decision thresholds are tuned on MIMIC-CXR validation and fixed for all evaluations.

TABLE XII
SENSITIVITY TO DISEASE-LABEL UNCERTAINTY HANDLING ON CHEXPERT PLUS EXPERT SUBSET (MACRO AUC).

Setting	Macro AUC
U-Ignore (default)	0.526
U-Zero	0.526

R. Full CheXpert-14 Results

Table **??** reports per-label AUC on all 14 CheXpert observations, complementing the main paper's focus on the five competition labels.

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

[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, 4

[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

[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

[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

[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] 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

[19] 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

[20] I. Kim, J. Kim, J. Choi, and H. J. Kim, "Concept bottleneck with visual concept filtering for explainable medical image classification," *arXiv preprint arXiv:2308.11920*, 2023. 1, 2

[21] W. Pang, X. Ke, S. Tsutsui, and B. Wen, "Integrating clinical knowledge into concept bottleneck models," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 243–253. 1, 2

[22] Y. Yang, M. Gandhi, Y. Wang, Y. Wu, M. S. Yao, C. Callison-Burch, J. C. Gee, and M. Yatskar, "A textbook remedy for domain shifts: Knowledge priors for medical image analysis," *arXiv preprint arXiv:2405.14839*, 2024. 1

[23] 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, 4

[24] 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. 1, 2, 4

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

[26] 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

[27] 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

[28] 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

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

[30] 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

[31] 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, 4

[32] D. Ma, J. Pang, M. B. Gotway, and J. Liang, "A fully open AI foundation model applied to chest radiography," *Nature*, vol. 643, no. 8071, pp. 488–498, 2025. 2

[33] 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. 2

[34] 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. 2, 4

[35] 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. 2

[36] 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

[37] M. Yuksekgonul, M. Wang, and J. Zou, "Post-hoc concept bottleneck models," in *International Conference on Learning Representations*, 2023. 2, 5

[38] 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

[39] K. P. Panousis, D. Ienco, and D. Marcos, "Coarse-to-fine concept bottleneck models," 2024. [Online]. Available: https://arxiv.org/abs/2310.02116 2, 5

[40] 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. 2

[41] N. Debole, P. Barbiero, F. Giannini, A. Passeggini, G. Ciravegna, and F. Precioso, "If concept bottlenecks are the question, are foundation models the answer?" *arXiv preprint arXiv:2504.19774*, 2025. 2

[42] S. Yan, W. K. Cheung, I. W. Tsang, K. Chiu, T. M. Tong, K. C. Cheung, and S. See, "AHIVE: Anatomy-aware hierarchical vision encoding for interactive radiology report retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 14 324–14 333. 2

[43] J.-B. Delbrouck, P. Chambon, Z. Chen, M. Varma, A. Johnston, L. Blankemeier, D. Van Veen, T. Bui, S. Truong, and C. Langlotz, "RadGraph-XL: A large-scale expert-annotated dataset for entity and relation extraction from radiology reports," in *Findings of the Association for Computational Linguistics: ACL 2024*. Bangkok, Thailand: Association for Computational Linguistics, 2024, pp. 12 902–12 915. 2

[44] S. Khanna, A. Dejl, K. Yoon, S. Q. Truong, H. Duong, A. Saenz, and P. Rajpurkar, "RadGraph2: Modeling disease progression in radiology reports via hierarchical information extraction," in *Proceedings of the 8th Machine Learning for Healthcare Conference*, ser. Proceedings of Machine Learning Research, vol. 219. PMLR, 2023, pp. 381–402. 2

[45] K. Donnelly, "Snomed-ct: The advanced terminology and coding system for ehealth," *Studies in Health Technology and Informatics*, vol. 121, pp. 279–290, 2006. 3

[46] H. Q. Nguyen, H. H. Pham, T. L. Le, M. Dao, K. Lam *et al.*, "Vindr-cxr: An open dataset of chest x-rays with radiologist annotations," PhysioNet, 2021, version 1.0.0. [Online]. Available: https://physionet.org/content/vindr-cxr/1.0.0/ 4

[47] Radiological Society of North America, "Rsna pneumonia detection challenge," Kaggle, 2018. [Online]. Available: https://www.kaggle.com/c/rsna-pneumonia-detection-challenge 4

[48] 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. 4

[49] E. Tiu, E. Talius, P. Patel, C. P. Langlotz, A. Y. Ng, and P. Rajpurkar, "Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning," *Nature Biomedical Engineering*, vol. 6, no. 12, pp. 1399–1406, 2022. 5

[50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2015. 5

[51] Y. Yang, A. Panagopoulou, S. Zhou, D. Jin, C. Callison-Burch, and M. Yatskar, "Language in a bottle: Language model guided concept bottlenecks for interpretable image classification," *arXiv preprint arXiv:2211.11158*, 2023. [Online]. Available: https://arxiv.org/abs/2211.11158 5

[52] 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. 5

[53] J. P. Cohen, J. D. Viviano, P. Bertin, P. Morrison, P. Torabian, M. Guarrera, M. P. Lungren, A. Chaudhari, R. Brooks, M. Hashir, and H. Bertrand, "TorchXRayVision: A library of chest X-ray datasets and models," in *Medical Imaging with Deep Learning*, 2022. [Online]. Available: https://github.com/mlmed/torchxrayvision 5

[54] N. Debole, P. Barbiero, F. Giannini, A. Passerini, S. Teso, and E. Marconato, "If concept bottlenecks are the question, are foundation models the answer?" *arXiv preprint arXiv:2504.19774*, 2025. [Online]. Available: https://arxiv.org/abs/2504.19774 5