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 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 translate that confidence into the kinds of criteria clinicians expect, such as measured ratios or other anatomically grounded evidence. 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 standardize terminology and relations [18], but they do not provide image-grounded labels, such as presence, laterality, or anatomical site, for individual radiographs, so the core supervision requirement remains unchanged. 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 have pursued two directions, neither 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], [21]. 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. The second integrates clinical knowledge by guiding models to prioritize clinically important concepts through alignment losses [22], [23]. However, such approaches require expert-provided importance rankings for each concept and have not been demonstrated to scale beyond small concept sets; enumeration-based importance weighting becomes intractable when dozens or hundreds of concepts are involved, as in chest radiography.

We pursue a different direction: repurposing existing clinical NLP tools as sources of concept supervision. Tools such as RadGraph [24] 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. 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 [25]. 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 [26]. A finding may be "present," "absent," "unchanged," or "cannot be excluded," distinctions that matter clinically and must be preserved in any derived supervision [27], [28]. 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 [29], [24], together with resources such as the Unified Medical Language System (UMLS) [30], 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 [31], [7], linking medical entities to ontologies such as UMLS [30], 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. In contrast to alignment-loss approaches that require per-concept importance annotations [22], RadCBM encodes clinical knowledge structurally: ontology grounding via UMLS provides semantic standardization, and the 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.

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, derives region abnormality targets by pooling RadGraph-extracted concept locations, and gates region-specific findings through those derived region scores (no separate region annotations), while constraining 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 [32]; CheXpert [7] and MIMIC-CXR [25] expanded scale while improving label quality and providing associated reports. Architectures from DenseNet-based CheXNet and CheXNeXt [6], [33] 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 [34] 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 [35], concept embeddings that relax strict bottlenecks [36], and interactive variants enabling test-time correction [37]. Applications span dermatology [38], ophthalmology [39], 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], [40]. Coarse-to-fine CBMs further introduce multilevel bottlenecks, tying coarse (global) concepts to fine (localized) concepts to capture low-level details while preserving interpretability [41]. Visual TCAV and related approaches refine concept scoring and selection [42], while GlanceNets [43] and concept-shift analyses [44] 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 [24], [24], 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 [31] 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 [28], and RadGraph extended extraction to full entity-relation graphs [24]. Assertion detection, distinguishing present, absent, and uncertain findings, remains critical, addressed by systems from NegEx [27] through modern neural classifiers [45]. 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 [30]. Neural linking methods, particularly SapBERT's self-alignment pretraining on UMLS synonyms [29], 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 [46] prompted medical adaptations: ConVIRT [47], MedCLIP [48], and BiomedCLIP [49] 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 [40].

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. Method

RadCBM predicts diagnoses through a two-stage concept bottleneck with anatomical gating (Fig. **??**). Stage 1 trains an 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 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 [24] 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 [29] and retrieve the nearest neighbor from an index of SNOMED-CT synonyms [18]. 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

$\mathcal{C}$ contains 1,312 ontology-grounded findings, two orders of magnitude larger than the 14-class vocabularies in standard benchmarks.

Mention masking and uncertainty. Unmentioned findings are not necessarily absent; radiologists document only what they deem relevant. We therefore construct a mention mask $m_{ik} \in \{0, 1\}$ indicating whether concept k was explicitly asserted in study i. Unmentioned concepts (m_{ik}=0) are excluded from the loss. For mentioned concepts, we derive targets $t_{ik} \in \{0, 0.5, 1\}$: absent assertions map to 0, present to 1, and uncertain to 0.5 with downweighted loss contribution.

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 $g : \mathcal{C} \rightarrow \mathcal{R}$. Region-level targets are obtained by max-pooling over constituent concepts:

$$\tilde{z}_{ir} = \max_{k:\, g(k)=r} t_{ik}. \tag{1}$$

Regions containing no mentioned concepts are masked from region supervision.

B. Model Architecture

Stage 1: Image to concepts and regions. A frozen pretrained vision encoder f_θ extracts image features. 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:

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

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:

$$\mathcal{L}_{\text{stage1}} = \underbrace{\frac{1}{\sum_{i,k} m_{ik}} \sum_{i,k} m_{ik} \cdot \text{BCE}(s_{ik}, t_{ik})}_{\mathcal{L}_{\text{concept}}} + \lambda_r\, \mathcal{L}_{\text{region}}, \tag{4}$$

where $\mathcal{L}_{\text{region}}$ is BCE between region logits and pooled targets $\tilde{z}_i$, masked for regions without mentioned concepts.

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:

$$\ell_{ij} = \sum_k W_{jk} \cdot \hat{c}_{ik}^{\text{gated}}. \tag{5}$$

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
$\hat{c}_k^{\text{gated}} \leftarrow \hat{z}_{g(k)} \cdot \hat{c}_k \quad \forall k$ ▷ Gated concepts
Minimize $\mathcal{L}_{\text{concept}} + \lambda_r \mathcal{L}_{\text{region}}$ over ϕ, W_r ▷ Mention-masked

Stage 2: Learn diagnosis head
Freeze ϕ, W_r
$\ell_j \leftarrow \sum_k W_{jk} \cdot \hat{c}_k^{\text{gated}} \quad \forall j$ ▷ 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 $\{W_{jk} \cdot \hat{c}_k^{\text{gated}}\}_{j,k}$

The contribution of concept k to diagnosis j is directly readable as $W_{jk} \cdot \hat{c}_{ik}^{\text{gated}}$: 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 -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 $W_{jk} \cdot \hat{c}_k^{\text{gated}}$ per diagnosis and threshold activations at δ=0.5 when summarizing.

Algorithm 1 summarizes the full procedure.

IV. Results

A. Experimental Setup

1) Datasets: We evaluate on five chest radiograph benchmarks spanning in-domain and external validation. **MIMIC-CXR** [25] 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** [50] provides radiologist annotations for 28 findings, **RSNA Pneumonia** [51] provides pneumonia detection labels with bounding boxes, and **NIH ChestX-ray14** [32] 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 [24], yielding **127,834** 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 (**XXX** concepts), heart (**XXX** concepts), pleura (**XXX** concepts), mediastinum (**XXX** concepts), bone (**XXX** concepts), and other (**XXX** concepts). Assertion status (present, absent, uncertain) is preserved for each concept mention.

3) Implementation Details: We implement RadCBM with frozen radiology-pretrained vision backbones, including MedCLIP [48], CXR-CLIP [52], and CheXzero [53]. 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 [54] with learning rate 10^{-4}, batch size 32, and early stopping based on validation macro AUC (patience 10 epochs). We set region loss weight $\lambda_r = 0.1$, temperature $\tau = 1.0$, and gate floor $\epsilon = 0.01$ based on validation performance; 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** [35], which retrofits concept bottlenecks onto pretrained models; (2) **LaBo** [55], which constructs text-defined bottlenecks with linear concept-to-class predictors; (3) **AdaCBM** [56], which adds an adaptive module to reduce domain mismatch; and (4) **C2F-CBM** [41], which builds two-level bottlenecks with coarse-to-fine prediction.

Black-box baselines. Supervised CNNs (ResNet-50, DenseNet-121) [57] 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 [58]:

- 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.

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).

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 matches or exceeds all CBM baselines while providing interpretable concept-mediated predictions. On MIMIC-CXR, RadCBM achieves a macro AUC of **0.XXX**, comparable to the supervised DenseNet-121 baseline (**0.XXX**) and outperforming all other CBM methods. Hierarchical gating improves over the flat variant by **X.X** points in macro AUC, with notable gains on region-specific pathologies such as Pleural Effusion (**+X.X**) and Edema (**+X.X**).

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. RadCBM achieves the highest overall concept AUC (**0.XXX**) and AUPRC (**0.XXX**). The improvement is pronounced for rare concepts: RadCBM attains **0.XXX** AUC on rare findings compared to **0.XXX** for the flat variant, consistent with hierarchical gating suppressing spurious activations when regions are predicted normal.

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. RESULTS ARE MEAN OVER 3 SEEDS; STD <0.01 OMITTED.

Method	Type	Atelect.	Cardiom.	Consolid.	Edema	Pl. Eff.	Macro
MIMIC-CXR Test Set							
ResNet-50	CNN	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**
DenseNet-121	CNN	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**
MedCLIP	VLM	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**
Post-hoc CBM	CBM	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**
LaBo	CBM	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**
AdaCBM	CBM	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**
C2F-CBM	H-CBM	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**
RadCBM (flat)	CBM	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**
RadCBM	H-CBM	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**
CheXpert Plus Expert Subset							
ResNet-50	CNN	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**
DenseNet-121	CNN	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**
MedCLIP	VLM	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**
Post-hoc CBM	CBM	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**
LaBo	CBM	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**
AdaCBM	CBM	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**
C2F-CBM	H-CBM	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**
RadCBM (flat)	CBM	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**
RadCBM	H-CBM	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**

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.

Method	AUC	AUPRC	Rare AUC	Rare AUPRC
Post-hoc CBM	**.XXX**	**.XXX**	**.XXX**	**.XXX**
LaBo†	**.XXX**	**.XXX**	**.XXX**	**.XXX**
AdaCBM	**.XXX**	**.XXX**	**.XXX**	**.XXX**
C2F-CBM	**.XXX**	**.XXX**	**.XXX**	**.XXX**
RadCBM (flat)	**.XXX**	**.XXX**	**.XXX**	**.XXX**
RadCBM	**.XXX**	**.XXX**	**.XXX**	**.XXX**

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.

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.

E. Interpretability

Table IV evaluates whether concept-based explanations support faithful interventions and clinically plausible activations. 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 **XX.X%** (flat) to **XX.X%**, indicating that region gates successfully suppress findings in anatomically inactive regions. Region consistency (**0.XXX**) confirms that gates align with pooled concept evidence.

F. Ablation Study

We report ablations isolating key design choices in the supplement (Table VI). Briefly: (1) assertion-aware mention masking improves concept AUC by **X.X** points by avoiding supervision corruption from negated findings; (2) hierarchical gating improves plausibility from **XX%** to **XX%** with minimal impact on classification AUC; (3) conservative soft-gating ($\epsilon > 0$) prevents cascading failures where missed region predictions suppress all constituent findings.

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	report	CheXpert-14
CheXpert Plus	radiologist-annotated	study	CheXpert-14
VinDr-CXR	radiologist-annotated	image	CheXpert-5
RSNA Pneumonia	radiologist-annotated	image+bbox	Pneumonia
NIH ChestX-ray14	report-derived	image	NIH-14

For MIMIC-CXR, test reports were annotated by a radiologist into CheXpert-14 categories; these are radiologist-annotated report labels rather than independent image readouts. 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:

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

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

TABLE III
EXTERNAL VALIDATION (AUC-ROC). MULTI-LABEL COLUMNS: MACRO AUC OVER FIVE CHEXPERT LABELS; RSNA: BINARY PNEUMONIA AUC.

Method	Type	MIMIC	CheXpert Plus	VinDr-CXR	NIH	RSNA
DenseNet-121	CNN	**.XXX**	**.XXX**	**.XXX**	**.XXX**	**.XXX**
MedCLIP	VLM	**.XXX**	**.XXX**	**.XXX**	**.XXX**	**.XXX**
LaBo	CBM	**.XXX**	**.XXX**	**.XXX**	**.XXX**	**.XXX**
AdaCBM	CBM	**.XXX**	**.XXX**	**.XXX**	**.XXX**	**.XXX**
C2F-CBM	H-CBM	**.XXX**	**.XXX**	**.XXX**	**.XXX**	**.XXX**
RadCBM (flat)	CBM	**.XXX**	**.XXX**	**.XXX**	**.XXX**	**.XXX**
RadCBM	H-CBM	**.XXX**	**.XXX**	**.XXX**	**.XXX**	**.XXX**

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†	**.XX**	**.XXX**	**.XXX**	**.XXX**
AdaCBM	**.XX**	**.XXX**	**.XXX**	**.XXX**
C2F-CBM	**.XX**	**.XXX**	**.XXX**	**.XXX**
RadCBM (flat)	**.XX**	**.XXX**	**.XXX**	—
RadCBM	**1.00**	**.XXX**	**.XXX**	**.XXX**

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 (**XXX** CONCEPTS) IS EXCLUDED AS IT AGGREGATES HETEROGENEOUS FINDINGS WITHOUT CLEAR ANATOMICAL LOCALIZATION.

Region	#Concepts	Region AUC	Finding AUC	Prevalence (%)
Lung	**XXX**	**.XXX±.XXX**	**.XXX±.XXX**	**XX.X**
Heart	**XXX**	**.XXX±.XXX**	**.XXX±.XXX**	**XX.X**
Pleura	**XXX**	**.XXX±.XXX**	**.XXX±.XXX**	**XX.X**
Mediastinum	**XXX**	**.XXX±.XXX**	**.XXX±.XXX**	**XX.X**
Bone	**XXX**	**.XXX±.XXX**	**.XXX±.XXX**	**XX.X**
Overall	1,312	**.XXX±.XXX**	**.XXX±.XXX**	—

E. Learned Concept-Label Relationships

Fig. 1. 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. 2. 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

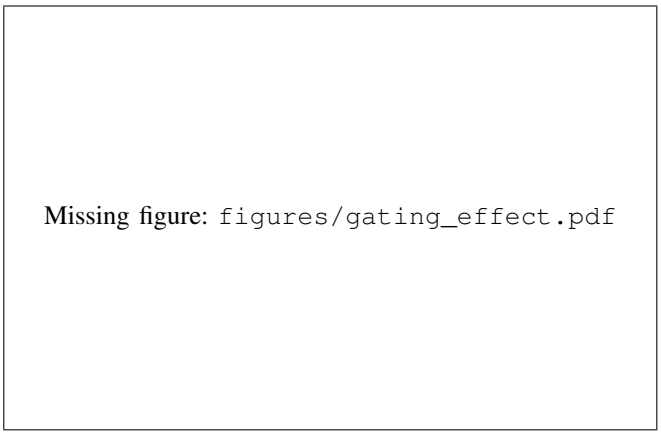

Fig. 3. Effect of hierarchical gating. (a) Region abnormality score versus mean finding activation for flat vs hierarchical variants. (b) Distribution of finding activations stratified by region gate status.

H. Intervention Faithfulness Analysis

Fig. 4. Intervention faithfulness. (a) Label probability as a function of concept activation. (b) Predicted versus observed label change upon concept intervention.

I. Hyperparameter Sensitivity

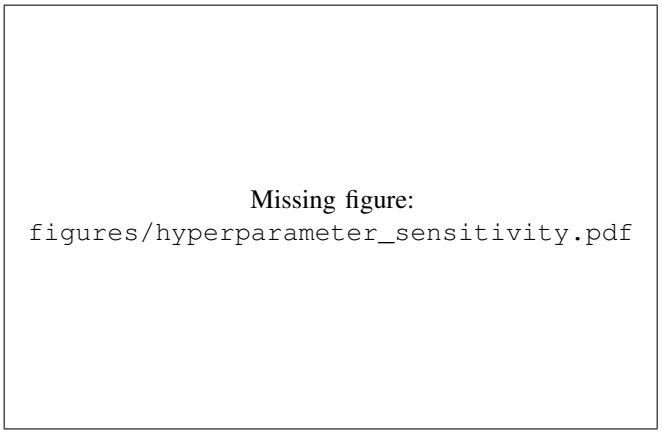

Fig. 5. Sensitivity to region loss weight λ_r. Validation macro AUC remains stable across $\lambda_r \in [0.01, 1.0]$.

J. Concept AUC by Frequency

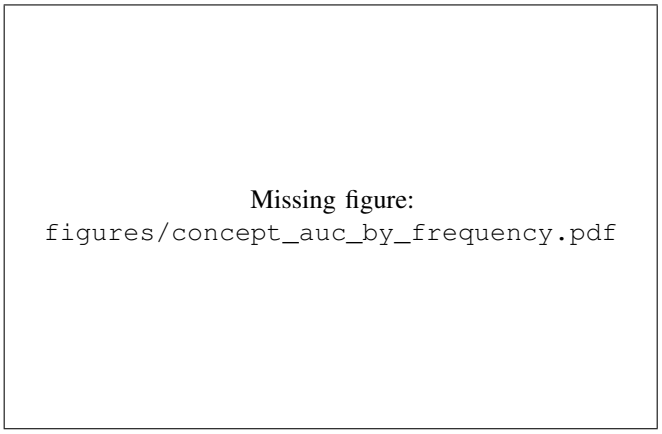

Fig. 6. Concept AUC stratified by training set frequency. Hierarchical gating provides larger gains for rare concepts.

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	**.XXX**	**−X.X%**
RadCBM	MIMIC → CheXpert	**.XXX**	**−X.X%**
DenseNet-121	CheXpert → MIMIC	**.XXX**	**−X.X%**
RadCBM	CheXpert → MIMIC	**.XXX**	**−X.X%**

M. Computational Requirements

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

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

N. Error Analysis

Missing figure: `figures/error_analysis.pdf`

Fig. 8. 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	**.XXX**	**.XXX**
MedCLIP	**.XXX**	**.XXX**
RadCBM	**.XXX**	**.XXX**

Missing figure:
`figures/reliability_diagram.pdf`

Fig. 9. Reliability diagram on CheXpert Plus expert subset.

P. Rare-Label Performance

TABLE XI
PR-AUC ON RARE LABELS (CHEXPERT PLUS EXPERT SUBSET).

Method	Fracture	Pneumothorax	Pneumonia	Lung Lesion	Pleural Other	Macro
DenseNet-121	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**
MedCLIP	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**
RadCBM	**.XX**	**.XX**	**.XX**	**.XX**	**.XX**	**.XXX**

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) with downweighted loss. Decision thresholds are tuned on MIMIC-CXR validation and fixed for all evaluations.

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, 3, 5

[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, 3

[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, 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

[21] 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

[22] 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

[23] 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

[24] 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, 5

[25] 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, 4

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

[27] 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

[28] 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

[29] 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

[30] 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

[31] 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

[32] 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, 5

[33] 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

[34] 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

[35] M. Yuksekgonul, M. Wang, and J. Zou, "Post-hoc concept bottleneck models," in *International Conference on Learning Representations*, 2023. 3, 5

[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. 3

[37] 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. 3

[38] 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. 3

[39] 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. 3

[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. 3

[41] K. P. Panousis, D. Ienco, and D. Marcos, "Coarse-to-fine concept bottleneck models," 2024. [Online]. Available: https://arxiv.org/abs/2310.02116 3, 5

[42] 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

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

[44] 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

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

[46] 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

[47] 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

[48] 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, 5

[49] 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

[50] 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/ 5

[51] Radiological Society of North America, "Rsna pneumonia detection challenge," Kaggle, 2018. [Online]. Available: https://www.kaggle.com/c/rsna-pneumonia-detection-challenge 5

[52] 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. 5

[53] 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

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

[55] 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

[56] 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

[57] 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

[58] 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