

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

Obadah Habash, Ahmed Alagha, Rabeb Mizouni, Shakti Singh, and Hadi Otrouk
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 demanding, 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

A. Concept Bank Construction from Reports

Let $\mathcal{D} = \{(x_i, r_i, y_i)\}_{i=1}^N$ denote a dataset of chest radiographs x_i , associated free-text reports r_i (when available), and multi-label targets $y_i \in \{0, 1, -1\}^L$ for L clinical labels (CheXpert-style), where -1 denotes uncertainty or missingness. RadCBM predicts labels through an explicit concept bottleneck: we distill a concept bank $\mathcal{C} = \{c_1, \dots, c_K\}$ from reports and train models that first predict concepts from images and then predict diagnoses from the predicted concepts, where $\hat{c} \in [0, 1]^K$ is a vector of concept probabilities.

a) *Entity extraction and assertion.*: We apply RadGraph-XL [24] to report findings and impression sections to extract entity mentions and their assertion status (*present*, *absent*, *uncertain*). For each mention we also retain modifier spans linked by RadGraph relations, which often encode anatomical context such as laterality or coarse location. These modifiers are normalized into a free-text location string that is carried forward as weak spatial context.

b) *Ontology linking.*: Each extracted mention is linked to a UMLS Concept Unique Identifier (CUI) using SapBERT [29] embeddings and nearest-neighbor retrieval over a synonym index built from UMLS. To reduce off-domain matches, we restrict candidate synonyms to a curated set of radiology-relevant semantic types and constrain sources to SNOMED CT terms [30], [18]. We embed both the mention surface form and the mention concatenated with its modifier tokens, then keep the highest-scoring candidate above a similarity threshold (default 0.8), discarding low-confidence links.

c) *Study-level aggregation and inventory.*: For each study, we aggregate linked mentions into a set of concept records containing the canonical concept name, linked CUI, assertion status, and derived location string. If a study contains multiple mentions mapping to the same concept, we keep the highest-precedence assertion, where present overrides uncertain and uncertain overrides absent, to avoid contradictory supervision when forming study-level labels. Aggregating across studies

yields an inventory with per-concept frequencies, supported assertions, and observed locations.

d) *Pruning and label tensors.*: To obtain a CBM-friendly vocabulary, we prune the inventory by concept category (findings by default) and frequency (for example, at least 10 total occurrences with at least one positive mention), with optional name-based filters to remove non-informative normality concepts. We then create an ordered concept index and a dense study-by-concept label matrix using a multi-assertion encoding $a_{ik} \in \{0, 1, 2, 3\}$ for (unmentioned, absent, uncertain, present). This encoding supports *mention-masked* supervision: unmentioned concepts are treated as unknown rather than negative. Although our pipeline can retain anatomy scaffold and device concepts, we use finding concepts by default to form the bottleneck representation. In the current pruned bank, this procedure yields $K = 1,312$ ontology-grounded findings.

e) *Optional MI-based filtering.*: When a more compact, task-specific concept set is desired, we optionally apply label-aware filtering using mutual information (MI) between concept presence and downstream labels. We binarize concept presence using only *present* assertions and compute $I(c_k; y_j)$ for each concept and label on the training set, then retain concepts that exceed a threshold or fall within the top K by $\max_j I(c_k; y_j)$. This step is optional and is used to trade vocabulary size for label relevance.

f) *Coarse anatomical regions.*: We define a coarse anatomical partition \mathcal{R} consisting of lung, pleura, heart, mediastinum, bone, and an additional other bucket. Each concept k is assigned to a parent region $g(k) \in \mathcal{R}$ using deterministic rules based on supported location strings, simple name cues (for example, pleur, pulmon, mediastin), and semantic type. This fixed concept to region map is shared across all hierarchical components.

Given any concept-valued vector $v \in [0, 1]^K$, we define pooled region scores by max-pooling over concepts in each region:

$$P_r(v) = \max_{k: g(k)=r} v_k \quad \forall r \in \mathcal{R}. \quad (1)$$

When report-derived concept supervision is available, we derive pooled region targets $\tilde{z}_{ir} = P_r(t_i)$ with a corresponding mask indicating whether any concept in region r is explicitly mentioned; unmentioned regions are excluded from region losses.

B. RadCBM Model and Training

RadCBM follows a two-stage design: an image→concept predictor trained from report-derived concept supervision, and a concept→label head trained to map predicted concepts to downstream diagnoses. Both stages can include region prediction and multiplicative gating through the shared concept to region map.

a) *Data alignment and splits.*: All training and evaluation are performed at the study level. We use predefined train, validation, and test splits and align (i) images, (ii) report-derived concept supervision, and (iii) disease-label CSV targets by study identifier to avoid leakage and to ensure that the

diagnosis head is trained on the same population for which concept predictions are produced.

b) *Concept supervision and uncertainty handling.*: From the multi-assertion matrix a_{ik} we derive targets $t_{ik} \in \{0, 0.5, 1\}$ and a mention mask $m_{ik} \in \{0, 1\}$. Unmentioned concepts are masked ($m_{ik} = 0$); absent and present are supervised as negative and positive; uncertain mentions are either ignored (masked) or treated as soft targets ($t_{ik} = 0.5$) with a reduced loss weight. For disease labels, we treat uncertain labels as missing by default and ignore them in the label loss and in per-class metrics, with optional mappings of uncertainty to positive or negative for sensitivity analyses.

c) *Stage 1: image to concepts (and regions).*: A radiology-pretrained vision backbone maps the image to a feature vector $h = f_\theta(x)$, implemented with MedCLIP [48] (ViT or ResNet variants) or standard ResNet backbones. A lightweight two-layer MLP produces concept logits $s = g_\phi(h)$ and probabilities $\hat{c} = \sigma(s)$. We optimize a mention-masked binary cross-entropy that supervises only concepts that are explicitly asserted:

$$\mathcal{L}_{\text{concept}} = \frac{1}{\sum_{i,k} m_{ik}} \sum_{i,k} m_{ik} \text{BCE}(s_{ik}, t_{ik}). \quad (2)$$

When a region map is available, we additionally learn a region head that predicts region logits from the intermediate concept activations. We supervise region logits using pooled targets from Eq. (1) when available, and we optionally clamp predicted region gates using these pooled targets, treating unknown regions as neutral. Region probabilities are used as multiplicative gates on concept activations, which discourages anatomically implausible concept predictions while preserving a soft failure mode controlled by the gate floor and temperature. The concept predictor can further be regularized with an ontology-aware graph Laplacian penalty over concept head weights:

$$\mathcal{L}_{\text{graph}} = \sum_{(p,q) \in E} \|w_p - w_q\|_2^2, \quad (3)$$

where E is a user-specified set of concept edges, including a simple within-region chain graph or externally provided edges. When region supervision is enabled, the Stage 1 objective becomes $\mathcal{L}_{\text{concept}} + \lambda_r \mathcal{L}_{\text{region}} + \lambda_g \mathcal{L}_{\text{graph}}$, where $\mathcal{L}_{\text{region}}$ is a mention-masked BCE between region logits and pooled region targets.

d) *Concept and region quality.*: We evaluate concept prediction quality using micro-averaged classification metrics and AUROC/AUPRC computed over the subset of concepts with non-masked targets. When region prediction is enabled, we analogously compute region metrics using pooled region targets derived from report supervision and report per-region concept quality by grouping concepts under $g(k)$.

e) *Stage 2: concepts to labels.*: After training the concept predictor, we export per-study concept probabilities and train a diagnosis head on these predicted concepts. We consider a flat CBM head (a small MLP) and a hierarchical region-gated head described next. The diagnosis head is trained with the concept predictor frozen, so performance reflects the quality of the learned concept bottleneck rather than end-to-end fine-tuning.

f) *Hierarchical region gating with linear label head.*: The hierarchical head predicts region logits $u = W_r \hat{c}$ and region probabilities

$$\hat{z} = \sigma\left(\frac{u}{\tau}\right), \quad \hat{z} \leftarrow \epsilon + (1 - \epsilon)\hat{z}, \quad (4)$$

where τ is a temperature and ϵ is an optional gate floor for conservative soft-gating. Each concept is gated by its parent region, yielding $\hat{c}_k^{\text{gated}} = \hat{z}_{g(k)} \hat{c}_k$. Labels are then predicted by a bias-free linear layer

$$\ell = W_y \hat{c}^{\text{gated}}. \quad (5)$$

This constraint makes explanations intrinsic: the signed contribution of concept k to label ℓ_j is $W_y[j, k] \hat{c}_k^{\text{gated}}$. Removing the bias term prevents the head from predicting a diagnosis in the absence of supporting concept evidence.

g) *Gate clamping and region consistency.*: During diagnosis-head training, we derive soft region targets by pooling the input concept vector, $P_r(\hat{c})$, and optionally clamp region gates by multiplying \hat{z} with these pooled targets. We also optionally include an auxiliary region loss that encourages region logits to match pooled targets. The full objective is

$$\mathcal{L}_{\text{CBM}} = \mathcal{L}_{\text{label}} + \lambda_r \mathcal{L}_{\text{region}}, \quad (6)$$

where $\mathcal{L}_{\text{label}}$ is a masked BCE on CheXpert-style labels and $\mathcal{L}_{\text{region}}$ is a BCE between u and pooled region targets. We train both stages with Adam, use early stopping based on validation performance, and export per-study concept probabilities (and region probabilities when enabled) for downstream training and inspection.

C. Interpretability and Concept Interventions

RadCBM produces structured intermediate outputs that support direct inspection. First, we report the predicted concept vector \hat{c} as an ontology-grounded explanation. Second, the region head produces region probabilities \hat{z} that summarize abnormality by coarse anatomy and control which findings can contribute to predictions. Third, for the hierarchical head, Eq. (4) yields intrinsic attributions via the per-concept contributions $W_y[j, k] \hat{c}_k^{\text{gated}}$.

a) *Anatomical plausibility.*: We quantify anatomical plausibility by measuring the rate of *implausible activations*, defined as cases where a concept is predicted present while its parent region gate is low. Operationally, for a threshold δ , an activation is implausible if $\hat{c}_k > \delta$ but $\hat{z}_{g(k)} < \delta$.

b) *Concept interventions.*: To assess intervention behavior, we edit selected concept entries in \hat{c} (setting \hat{c}_k to 0 or 1) and measure the induced change in predicted label probabilities. In intrinsic CBMs, the signed contributions in Eq. (4) provide a testable prediction for the direction and relative magnitude of these counterfactual effects.

c) *Gate plausibility and intervention summaries.*: In addition to standard label metrics, we summarize gate plausibility by aggregating implausible activation rates across concepts and across regions. For trained hierarchical heads, we also report compact intervention summaries by ranking concepts by the magnitude of their learned contributions and estimating the expected change in label logits under concept edits.

IV. RESULTS

A. Experimental Setup

1) *Datasets.*: We evaluate on five chest radiograph benchmarks spanning in-domain testing 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/diagnoses, **RSNA Pneumonia** [51] provides pneumonia detection labels (and bounding boxes for positive cases), and **NIH ChestX-ray14** [32] provides 14 disease labels originally mined from reports. In addition, the test subset of the MIMIC-CXR radiology reports were annotated by a single radiologist into one of fourteen categories (CheXpert-14 style, including “No Finding”). CheXpert Plus (expert subset), VinDr-CXR, and RSNA Pneumonia provide radiologist-annotated evaluation labels, while NIH ChestX-ray14 is primarily report/NLP-derived for standard disease labels. We therefore emphasize performance on radiologist-labeled evaluation subsets as primary evidence of clinical correctness, and treat purely report/NLP-derived targets as complementary large-scale evidence.

For report-bearing datasets, we obtain 14-label targets using the CheXpert labeler and, to reduce label noise, we also consider an ensemble of complementary labelers (CheXpert, CheXbert [28], and NegBio [31]). The ensemble maps labeler outputs to {positive, negative, uncertain}. To avoid overstating performance on purely NLP-derived targets, we emphasize results on radiologist-labeled evaluation subsets (CheXpert Plus expert subset; and, when available, VinDr-CXR/RSNA Pneumonia) as primary evidence of clinical correctness.

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 five anatomical regions: lung (**142** concepts), heart (**38** concepts), pleura (**47** concepts), mediastinum (**51** concepts), and bone (**34** concepts). Assertion status (present, absent, uncertain) is preserved for each concept mention.

3) *Implementation Details.*: We implement RadCBM with interchangeable radiology-pretrained vision backbones, and report results for multiple backbones to assess robustness to representation choice. Specifically, we consider radiology-pretrained encoders including MedCLIP [48], CXR-CLIP [52], and CheXzero [53]. Unless otherwise stated, all concept-based methods within a comparison share the same backbone and are trained with identical optimization settings.

Images are resized to the backbone’s input resolution and normalized using the corresponding preprocessing. We apply standard augmentations during training: random horizontal flipping, rotation ($\pm 10^\circ$), 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 with patience of 10 epochs. Loss weights are set to $\lambda_1 = 0.5$ and $\lambda_2 = 1.0$ based

on validation performance. All experiments were conducted on an Intel i7-11800H @ 2.30GHz workstation equipped with 64GB RAM and an NVIDIA GeForce RTX 3080 GPU (16GB VRAM) using PyTorch 2.0. To ensure statistical reliability, we report results averaged over 3 random seeds with different weight initializations.

4) *Baselines*: Our benchmarks fall into two categories: concept bottleneck models (CBMs) and black-box baselines.

CBM benchmarks: (1) **Post-hoc CBM** [35], which retrofits concept bottlenecks onto pretrained models; (2) **Language-Guided Bottlenecks (LaBo)** [55], which constructs a text-defined bottleneck and a linear concept-to-class predictor; (3) **AdaCBM** [56], which adds an adaptive module between CLIP features and the bottleneck to reduce source–target mismatch; and (4) **Coarse-to-Fine CBM (C2F-CBM)** [41], which builds a two-level bottleneck by predicting coarse concepts from global image features and fine concepts from localized (patch/region) evidence with optional hierarchical tying. In our setting, the coarse level corresponds to anatomical region abnormality and the fine level corresponds to region-specific findings from our concept bank; fine predictions are aggregated across patches and tied to regions through the region–finding hierarchy.

Shared concept bank for concept-level evaluation: to make concept-level evaluations comparable across CBMs, we evaluate CBM concept predictors (including LaBo, AdaCBM, C2F-CBM, and RadCBM) on the same 1,312 ontology-grounded region-specific concepts extracted from MIMIC-CXR training reports (RadGraph+UMLS). Following recent recommendations for fair evaluation of VLM-CBMs with a fixed “gold” concept vocabulary [57], we use this shared concept bank as the concept target set for concept quality (Table II). For intervention-based interpretability metrics (Table IV), we restrict comparisons to *intrinsic* CBMs where the label prediction is mediated by the bottleneck (post-hoc CBMs are excluded since intervening on auxiliary concepts does not change the underlying predictor). For LaBo we use an ontology-aligned variant (denoted “LaBo (fixed vocab)”) that takes our concept bank as the candidate pool. We still report LaBo’s standard setup for label performance (Tables I and III).

Black-box models: we include (1) supervised CNN baselines (**ResNet-50**, **DenseNet-121**) [58]; and (2) vision-language models evaluated as black-box vision encoders, including **MedCLIP** [48], **CXR-CLIP** [52], and **CheXzero** [53].

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; threshold-free). Full CheXpert-14 results are reported in the supplementary material. When thresholded metrics (e.g., F1) are reported, we tune *per-label* decision thresholds on the official MIMIC-CXR validation split and keep these thresholds fixed for MIMIC-CXR test and all external benchmarks. **Concept quality** is assessed on the shared 1,312-concept bank using macro AUC-ROC and macro AUPRC (macro-AP), reported overall and on a rare-concept subset. **Interpretability** (intrinsic CBMs only) is evaluated via: (1) *Intervention faithfulness*, the Pearson correlation

between predicted concept contribution ($w_i \cdot c_i$) and observed label change upon concept intervention; (2) *Plausibility*, the fraction of activated findings ($c_i > 0.5$) whose parent region abnormality exceeds 0.5; (3) *Implausible activation rate*, the fraction of finding activations occurring when the parent region score is below 0.3. Unless otherwise stated, all reported numbers are mean \pm standard deviation over 3 random seeds.

B. Classification Performance

Table I presents classification performance across CBM and black-box benchmarks on the five CheXpert competition labels. RadCBM matches or exceeds the strongest CBM baselines while providing interpretable concept-mediated predictions. On MIMIC-CXR, RadCBM attains a macro AUC of **0.XXX**, matching a strong supervised baseline (**0.XXX**) and outperforming all CBM baselines. The hierarchical architecture improves over the flat variant by **X.X** percentage points in macro AUC, with notable gains on region-specific pathologies such as Pleural Effusion (**+X.X%**) and Edema (**+X.X%**).

Vision-language models treated as black-box vision encoders (MedCLIP, CXR-CLIP, CheXzero) achieve reasonable zero-shot performance but fall short of supervised CNNs and CBMs, particularly for rare findings. Among CBM approaches, methods relying on small concept vocabularies or automatically generated concepts tend to exhibit lower classification performance, suggesting that ontology-grounded concept banks with broader coverage provide stronger supervisory signal.

We next report concept quality, external validation on radiologist-annotated benchmarks, and interpretability-focused analyses.

C. Concept Quality

Table II compares concept prediction quality across methods on a shared target set of 1,312 ontology-grounded concepts. Since concepts are highly imbalanced, we report both macro AUC-ROC and macro AUPRC (macro-AP), overall and on rare concepts (50–200 training occurrences). RadCBM achieves the highest overall concept AUC (**0.XXX**) and macro AUPRC (**0.XXX**), outperforming other CBM baselines using the same concept bank. The improvement is particularly pronounced for rare concepts: RadCBM attains **0.XXX** AUC and **0.XXX** macro AUPRC on this subset compared to **0.XXX/0.XXX** for the flat variant, consistent with hierarchical gating suppressing implausible activations when regions are predicted normal.

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

D. External Validation on Radiologist-Annotated Benchmarks

To mitigate over-reliance on NLP-derived targets, we report external validation on benchmarks with radiologist-annotated evaluation labels when available (Table III). For multi-label

TABLE I

CLASSIFICATION PERFORMANCE (AUC-ROC) ON THE MIMIC-CXR TEST SET AND THE CHEXPRT PLUS VALIDATION SET (EXPERT-LABELED SUBSET) FOR THE FIVE CHEXPRT COMPETITION LABELS (ATELECTASIS, CARDIOMEGALY, CONSOLIDATION, EDEMA, PLEURAL EFFUSION). BEST RESULTS IN **BOLD**, SECOND-BEST UNDERLINED. CNN: SUPERVISED CNN BASELINE [58]; VLM: VISION-LANGUAGE MODEL (BLACK-BOX VISION ENCODER); CBM: CONCEPT BOTTLENECK MODEL; H-CBM: HIERARCHICAL CBM. ALL CONCEPT-BASED METHODS SHARE THE SAME VISUAL BACKBONE WITHIN EACH COMPARISON. RESULTS ARE MEAN OVER 3 SEEDS; STANDARD DEVIATIONS <0.01 OMITTED FOR CLARITY.

Method	Type	Atelectasis	Cardiomegaly	Consolidation	Edema	Pleural Eff.	Macro
<i>MIMIC-CXR Test Set</i>							
ResNet-50	CNN	.XX	.XX	.XX	.XX	.XX	.XXX
DenseNet-121	CNN	.XX	.XX	.XX	.XX	.XX	.XXX
MedCLIP (ViT)	VLM	.XX	.XX	.XX	.XX	.XX	.XXX
CXR-CLIP (ViT)	VLM	.XX	.XX	.XX	.XX	.XX	.XXX
CheXzero (SwinTiny)	VLM	.XX	.XX	.XX	.XX	.XX	.XXX
Post-hoc CBM	CBM	.XX	.XX	.XX	.XX	.XX	.XXX
LaBo CBM	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 (hier.)	H-CBM	.XX	.XX	.XX	.XX	.XX	.XXX
<i>CheXpert Plus Validation Set</i>							
ResNet-50	CNN	.XX	.XX	.XX	.XX	.XX	.XXX
DenseNet-121	CNN	.XX	.XX	.XX	.XX	.XX	.XXX
MedCLIP (ViT)	VLM	.XX	.XX	.XX	.XX	.XX	.XXX
CXR-CLIP (ViT)	VLM	.XX	.XX	.XX	.XX	.XX	.XXX
CheXzero (SwinTiny)	VLM	.XX	.XX	.XX	.XX	.XX	.XXX
Post-hoc CBM	CBM	.XX	.XX	.XX	.XX	.XX	.XXX
LaBo CBM	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 (hier.)	H-CBM	.XX	.XX	.XX	.XX	.XX	.XXX

datasets we report macro AUC-ROC over the five CheXpert competition labels (Atelectasis, Cardiomegaly, Consolidation, Edema, Pleural Effusion) using dataset-specific mappings (e.g., NIH ChestX-ray14 “Effusion” \leftrightarrow Pleural Effusion). For RSNA Pneumonia we report binary pneumonia AUC-ROC. For thresholded metrics (reported in supplementary when applicable), we use the same per-label thresholds tuned on MIMIC-CXR validation and do not tune on any external dataset.

E. Interpretability and Faithfulness

Table IV evaluates whether concept-based explanations support predictable interventions and clinically plausible activations. Intervention faithfulness measures whether predicted concept contributions $w_i \cdot c_i$ match observed label changes under concept editing; plausibility and implausible activation rate quantify whether findings activate primarily when their parent region is abnormal. We report these intervention-based metrics only for intrinsic CBMs where the bottleneck mediates the label prediction (post-hoc CBMs are not included), and since these models are evaluated using the same region-specific concept bank, the metrics are directly comparable across Table IV.

F. Ablations and Robustness (Supplementary)

We report a compact implementation ablation in the supplementary material (Table VI) that incrementally adds label cleanup (assertion-aware mention masking and labeler ensemble), conservative soft-gating, and ontology-aware regularization. This isolates which components drive label performance versus which primarily improve clinical plausibility and intervention faithfulness.

G. Summary

In the main paper, we emphasize (i) diagnostic performance on MIMIC-CXR and the radiologist-labeled CheXpert Plus expert subset, (ii) concept quality at scale (1,312 ontology-grounded concepts), (iii) faithful and clinically plausible interventions enabled by hierarchical gating, and (iv) robustness ablations for uncertainty handling and assertion-aware supervision. Additional analyses (region-level breakdowns, qualitative case studies, hyperparameter sensitivity, compute, and error analysis) are reported in the supplementary material.

APPENDIX A SUPPLEMENTARY MATERIAL

A. Evaluation Label Provenance

Table V summarizes which benchmarks provide radiologist-annotated evaluation labels versus report/NLP-derived labels for standard disease targets.

TABLE V

EVALUATION LABEL SOURCE (AT TEST TIME). WE REPORT WHETHER EVALUATION LABELS ARE RADIOLOGIST-ANNOTATED OR REPORT-DERIVED, ALONG WITH THEIR GRANULARITY AND THE LABEL SET USED IN OUR EVALUATION PROTOCOL.

Benchmark	Eval Label Source	Label Level	Eval Label Set
MIMIC-CXR (test reports)	radiologist-annotated	report-level	CheXpert-14
CheXpert Plus (expert subset)	radiologist-annotated	study-level	CheXpert-14
VinDr-CXR	radiologist-annotated	image-level	CheXpert-5
RSNA Pneumonia	radiologist-annotated	bbox+image-level	Pneumonia
NIH ChestX-ray14	report-derived	image-level	NIH14 (5 overlap)

For MIMIC-CXR, the test subset of the radiology reports were annotated by a single radiologist into one of fourteen categories (CheXpert-14 style); these are radiologist-annotated *report* labels rather than independent image readouts. NIH ChestX-ray14 evaluation labels are report/NLP-derived; we therefore treat them as complementary large-scale evidence and emphasize radiologist-labeled evaluation subsets (CheXpert Plus expert subset, VinDr-CXR, RSNA Pneumonia) as primary evidence of clinical correctness.

TABLE II

CONCEPT PREDICTION QUALITY ON MIMIC-CXR TEST SET FOR A SHARED SET OF 1,312 ONTOLOGY-GROUNDED CONCEPTS (RADGRAPH+UMLS). WE REPORT MACRO AUC-ROC AND MACRO AUPRC ACROSS CONCEPTS, OVERALL AND ON RARE CONCEPTS (50–200 TRAINING OCCURRENCES). [†]LaBo (FIXED VOCAB) DENOTES THE ONTOLOGY-ALIGNED VARIANT USED FOR CONCEPT-LEVEL EVALUATION WITH THE SHARED CONCEPT BANK. RESULTS AVERAGED OVER 3 SEEDS; \pm INDICATES STANDARD DEVIATION.

Method	AUC \uparrow	AUPRC \uparrow	Rare AUC \uparrow	Rare AUPRC \uparrow
Post-hoc CBM	.XXX \pm .XXX	.XXX \pm .XXX	.XXX \pm .XXX	.XXX \pm .XXX
LaBo (fixed vocab) [†]	.XXX \pm .XXX	.XXX \pm .XXX	.XXX \pm .XXX	.XXX \pm .XXX
AdaCBM	.XXX \pm .XXX	.XXX \pm .XXX	.XXX \pm .XXX	.XXX \pm .XXX
C2F-CBM	.XXX \pm .XXX	.XXX \pm .XXX	.XXX \pm .XXX	.XXX \pm .XXX
RadCBM (flat)	.XXX \pm .XXX	.XXX \pm .XXX	.XXX \pm .XXX	.XXX \pm .XXX
RadCBM (hier.)	.XXX \pm .XXX	.XXX \pm .XXX	.XXX \pm .XXX	.XXX \pm .XXX

TABLE III

EXTERNAL VALIDATION ACROSS BENCHMARKS (AUC-ROC). MULTI-LABEL COLUMNS REPORT MACRO AUC-ROC OVER THE FIVE CHEXPert COMPETITION LABELS USING DATASET-SPECIFIC LABEL MAPPINGS (NOTE THAT VINDr-CXR AND NIH CHESTX-RAY14 DEFINE LABEL SETS THAT DIFFER FROM CHEXPert-14); RSNA PNEUMONIA REPORTS BINARY PNEUMONIA AUC-ROC. CHEXPert IS EVALUATED ON THE EXPERT-LABELED CHEXPert PLUS SUBSET.

Method	Type	MIMIC (5)	CheXpert Plus (5)	VinDr-CXR (5)	NIH (5)	RSNA (Pneumonia)
ResNet-50	CNN	.XXX	.XXX	.XXX	.XXX	.XXX
DenseNet-121	CNN	.XXX	.XXX	.XXX	.XXX	.XXX
MedCLIP (ViT)	VLM	.XXX	.XXX	.XXX	.XXX	.XXX
CXR-CLIP (ViT)	VLM	.XXX	.XXX	.XXX	.XXX	.XXX
CheXzero (SwinTiny)	VLM	.XXX	.XXX	.XXX	.XXX	.XXX
Post-hoc CBM	CBM	.XXX	.XXX	.XXX	.XXX	.XXX
LaBo CBM	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 (hier.)	H-CBM	.XXX	.XXX	.XXX	.XXX	.XXX

TABLE IV

INTERPRETABILITY METRICS ON MIMIC-CXR TEST SET (INTRINSIC CBMs ONLY). INTERVENTION FAITHFULNESS MEASURES CORRELATION BETWEEN PREDICTED AND OBSERVED LABEL CHANGES UPON CONCEPT EDITING. PLAUSIBILITY AND IMPLAUSIBLE ACTIVATION RATE QUANTIFY ALIGNMENT BETWEEN FINDING ACTIVATIONS AND REGION PREDICTIONS. [†]LaBo (FIXED VOCAB) DENOTES THE ONTOLOGY-ALIGNED VARIANT USED FOR CONCEPT-LEVEL EVALUATION WITH THE SHARED CONCEPT BANK. RESULTS AVERAGED OVER 3 SEEDS; \pm INDICATES STANDARD DEVIATION.

Method	Intervention Faithfulness \uparrow	Plausibility \uparrow	Implausible Act. Rate \downarrow	Region Consistency \uparrow
LaBo (fixed vocab) [†]	.XX \pm .XX	.XXX \pm .XXX	.XXX \pm .XXX	.XXX \pm .XXX
AdaCBM	.XX \pm .XX	.XXX \pm .XXX	.XXX \pm .XXX	.XXX \pm .XXX
C2F-CBM	.XX \pm .XX	.XXX \pm .XXX	.XXX \pm .XXX	.XXX \pm .XXX
RadCBM (flat)	.XX \pm .XX	.XXX \pm .XXX	.XXX \pm .XXX	—
RadCBM (hier.)	.XX \pm .XX	.XXX \pm .XXX	.XXX \pm .XXX	.XXX \pm .XXX

B. Ablation Study

TABLE VI

COMPACT IMPLEMENTATION ABLATION ON MIMIC-CXR TEST SET. WE INCREMENTALLY ADD COMPONENTS WHILE KEEPING THE EVALUATION PROTOCOL FIXED (THRESHOLDS TUNED ON MIMIC VALIDATION AND THEN FROZEN). RESULTS AVERAGED OVER 3 SEEDS.

Configuration	Macro AUC	Concept AUC	Region-level Plausibility	Region-level Faith.
RadCBM (base)	.XXX	.XXX	.XXX	.XXX
+ Label cleanup (mask+assertion + labeler ensemble)	.XXX	.XXX	.XXX	.XXX
+ Conservative soft-gating	.XXX	.XXX	.XXX	.XXX
+ Ontology-aware regularization	.XXX	.XXX	.XXX	.XXX
RadCBM (full)	.XXX	.XXX	.XXX	.XXX

C. Impact of Assertion Modeling

Assertion-aware mention masking is included in the label-cleanup row of Table VI. This change is most impactful for frequently negated findings (e.g., “no effusion”), since treating negated or unmentioned concepts as negative can corrupt supervision and inflate spurious activations.

Table VII reports performance decomposed by anatomical region on MIMIC-CXR. Region abnormality AUC is computed from pooled concept locations (not standalone region labels) using surrogate region targets obtained by max-pooling present concepts per region from RadGraph outputs.

TABLE VII
REGION-LEVEL PERFORMANCE ON MIMIC-CXR TEST SET. REGION AUC (REGION ABNORMALITY AUC; COMPUTED FROM POOLED CONCEPT LOCATIONS, NOT STANDALONE REGION LABELS) MEASURES BINARY ABNORMALITY DETECTION; FINDING AUC MEASURES CONCEPT PREDICTION WITHIN EACH REGION. RESULTS AVERAGED OVER 3 SEEDS.

Region	#Concepts	Region AUC	Finding AUC	Prevalence (%)
Lung	142	.XXX±.XXX	.XXX±.XXX	XX.X
Heart	38	.XXX±.XXX	.XXX±.XXX	XX.X
Pleura	47	.XXX±.XXX	.XXX±.XXX	XX.X
Mediastinum	51	.XXX±.XXX	.XXX±.XXX	XX.X
Bone	34	.XXX±.XXX	.XXX±.XXX	XX.X
Overall	1,312	.XXX±.XXX	.XXX±.XXX	—

E. Learned Concept–Label Relationships

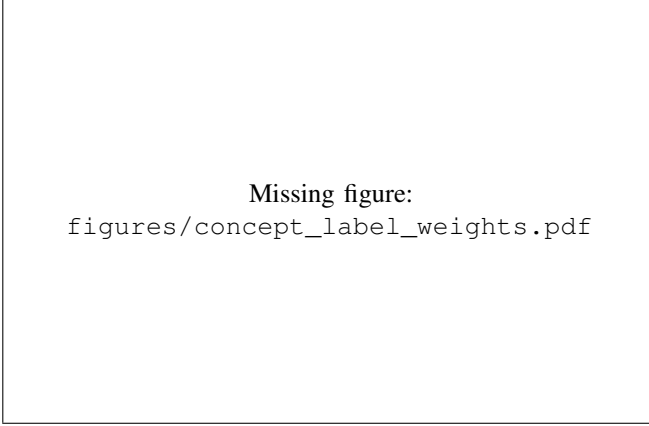


Fig. 1. Learned concept-to-label weights from the linear head. Each row shows the top-5 positive and top-5 negative concept contributions for one CheXpert label.

F. Concept Bank Analysis

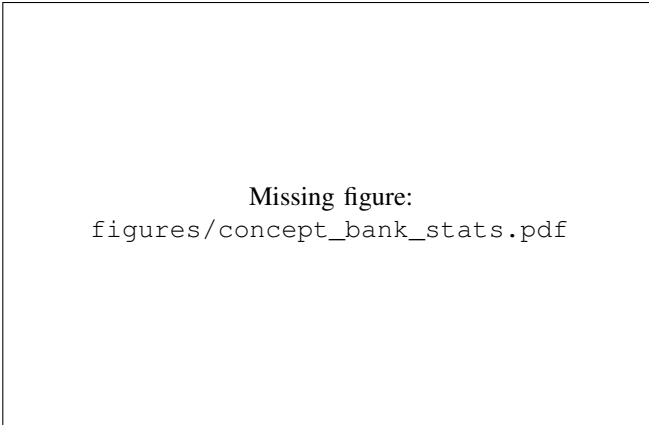


Fig. 2. Concept bank statistics. (a) Concept frequency distribution on log scale; vertical lines indicate CheXpert-14 concept positions. (b) Hierarchical organization by anatomical region; segment size proportional to concept count. (c) Vocabulary coverage comparison.

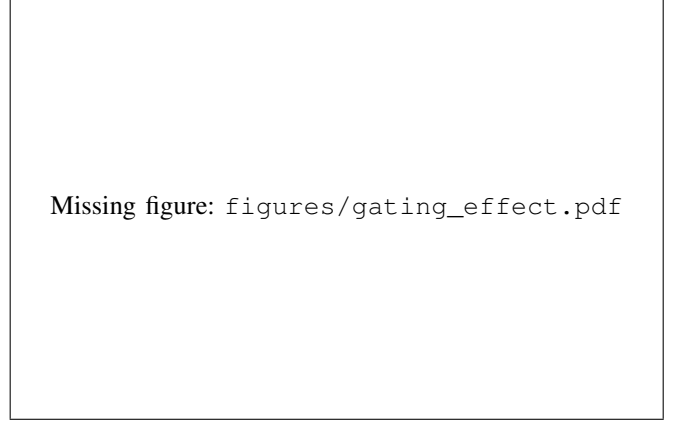


Fig. 3. Effect of hierarchical gating on concept activations. (a) Region abnormality score versus mean finding activation for flat vs hierarchical CBM. (b) Distribution of finding activations stratified by region status.

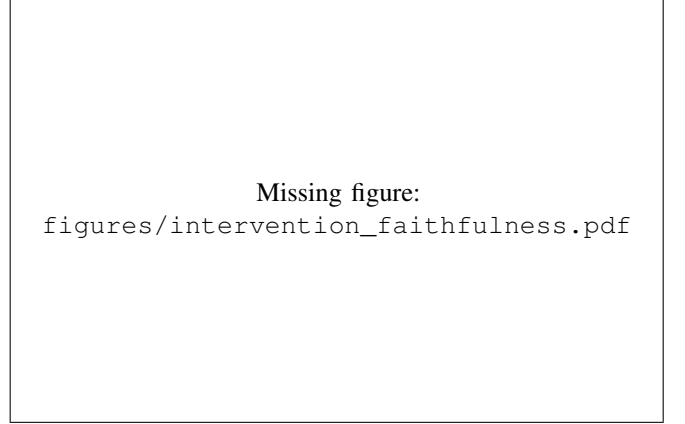


Fig. 4. Intervention faithfulness analysis. (a) Label probability as a function of concept activation. (b) Predicted concept contribution versus observed label change upon intervention.

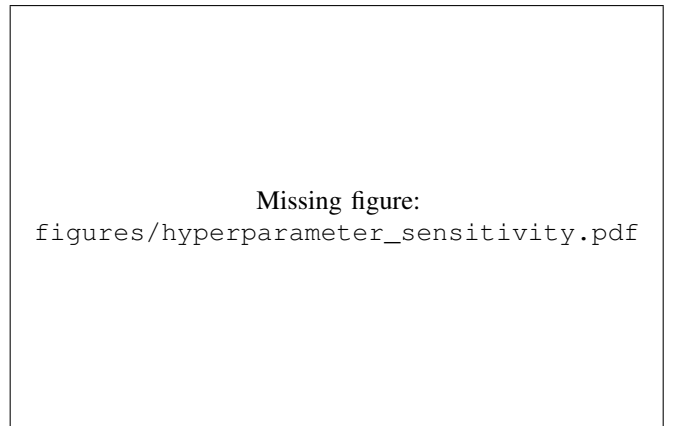


Fig. 5. Hyperparameter sensitivity analysis. Heatmap shows validation macro AUC across loss weight combinations (λ_1, λ_2).

Missing figure:
figures/concept_auc_by_frequency.pdf

Fig. 6. Concept AUC stratified by training set frequency.

G. Effect of Hierarchical Gating

H. Intervention Faithfulness Curves

I. Hyperparameter Sensitivity

J. Concept AUC by Frequency

K. Qualitative Case Studies

L. Cross-Dataset Generalization

TABLE VIII

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

Method	Train \rightarrow Test	Macro AUC	Δ from In-Domain
DenseNet-121	MIMIC \rightarrow CheXpert	.XXX	−X.X%
RadCBM (hier.)	MIMIC \rightarrow CheXpert	.XXX	−X.X%
DenseNet-121	CheXpert \rightarrow MIMIC	.XXX	−X.X%
RadCBM (hier.)	CheXpert \rightarrow MIMIC	.XXX	−X.X%

M. Computational Efficiency

TABLE IX

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

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

N. Error Analysis

O. Calibration

We report calibration on radiologist-annotated evaluation splits (e.g., CheXpert Plus expert subset, VinDr-CXR, and/or RSNA Pneumonia) using expected calibration error (ECE), Brier score, and reliability diagrams.

P. Rare-Label Performance (PR-AUC)

To complement ROC-AUC on imbalanced labels, we report PR-AUC (average precision) per label, emphasizing rare findings.

TABLE X
CALIBRATION METRICS. ECE AND BRIER SCORE COMPUTED ON CHEXPert PLUS EXPERT SUBSET; LOWER IS BETTER.

Method	ECE \downarrow	Brier \downarrow
DenseNet-121	.XXX	.XXX
MedCLIP	.XXX	.XXX
RadCBM (hier.)	.XXX	.XXX

TABLE XI
PER-LABEL PR-AUC ON 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 (hier.)	.XX	.XX	.XX	.XX	.XX	.XXX

Q. Protocol Notes: Uncertain Labels and Labeler Ensemble

Unless otherwise stated, we map labeler outputs to {positive, negative, uncertain}. When using an ensemble of report labelers (CheXpert, CheXbert, NegBio), disagreements are marked uncertain to reduce noise. We tune decision thresholds on the MIMIC-CXR validation split and report mean performance over 3 seeds.

R. Full CheXpert-14 Classification Results

Table XII reports full per-label AUC-ROC results on all 14 CheXpert observations for MIMIC-CXR and CheXpert Plus. These results complement the main-text evaluation, which focuses on the five CheXpert competition labels commonly used by MedCLIP and CheXzero.

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. Haghighi, 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

Missing figure: figures/qualitative_cases.pdf

Fig. 7. Qualitative case studies illustrating region-aware explanations.

TABLE XII

FULL CHEXPART-14 CLASSIFICATION PERFORMANCE (AUC-ROC) ON THE MIMIC-CXR TEST SET AND THE CHEXPART PLUS VALIDATION SET. BEST RESULTS IN **BOLD**, SECOND-BEST UNDERLINED. CNN: SUPERVISED CNN BASELINE [58]; VLM: VISION-LANGUAGE MODEL (BLACK-BOX VISION ENCODER); CBM: CONCEPT BOTTLENECK MODEL; H-CBM: HIERARCHICAL CBM. ALL CONCEPT-BASED METHODS SHARE THE SAME VISUAL BACKBONE WITHIN EACH COMPARISON. RESULTS AVERAGED OVER 3 SEEDS; STANDARD DEVIATIONS <0.01 OMITTED FOR CLARITY.

[illegible]

Missing figure: figures/error_analysis.pdf

Fig. 8. Error analysis. (a) Region-level confusion matrix showing prediction errors. (b) False-negative cascade: missed findings due to incorrect region normality prediction.

Missing figure:
figures/reliability_diagram.pdf

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

- [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. Shamsad, 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-Kobligh, 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, 4
- [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, 4, 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, 5
- [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, 5
- [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, 4
- [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, 4
- [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, 5
- [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, 4, 5, 6
- [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, 6
- [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, 6
- [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] 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> 6
- [58] 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, "TorchXRyVision: A library of chest X-ray datasets and models," in *Medical Imaging with Deep Learning*, 2022. [Online]. Available: <https://github.com/mlmed/torchxrayvision> 6, 12