

ARTICLE

Disease-Aware Multi-Relational Representation Learning and Neighborhood-Prior Decision Calibration for Chest X-Ray Multi-Label Prediction

Yuxin Zhang^{1, #}, Bin Li^{1, *, #}, Riqiang Liao² and Lianfang Tian¹

¹School of Automation Science and Engineering, South China University of Technology, Guangzhou, China

²Guangdong Lung Cancer Institute, Guangdong Provincial People's Hospital & Guangdong Academy of Medical Sciences, Guangzhou, China

*Corresponding Author: Bin Li. Email: binlee@scut.edu.cn

#These authors contributed equally to this work

Received: 24 June 2026; Accepted: 01 September 2026; Published: 28 September 2026

ABSTRACT: Chest X-ray multi-label prediction aims to identify multiple thoracic findings from radiographs and generate structured disease information for clinical analysis. Conventional image-only supervised learning uses disease labels primarily as prediction targets and may underutilize the cross-sample relational structure contained in multi-label disease annotations, thereby limiting disease-aware representation learning. Structured disease-label-guided contrastive pre-training provides a promising solution by encoding report-derived disease-label vectors as an auxiliary training view and aligning them with image representations through contrastive learning, while preserving image-only inference. In this formulation, the binary disease-label vector serves as both the structured input to the auxiliary encoder and the downstream prediction target, allowing its relational structure to guide image representation learning without introducing additional patient-level clinical variables. However, two major challenges remain. First, conventional instance-level contrastive alignment treats only the matched image and disease-label representation as a positive pair and regards other samples as negatives, which may introduce false-negative supervision when different patients share related disease combinations or label co-occurrence patterns. Second, fixed-threshold decision rules ignore label imbalance and patient-specific disease tendencies when converting continuous probabilities into binary labels. To address these limitations, we propose a two-stage framework for chest X-ray multi-label prediction. During pre-training, Multi-Relational Cross-Modal Alignment (MRCA) constructs clinical multi-relation targets from the paired identity relation, disease-label similarity, and tabular phenotype similarity, and converts them into soft target distributions for bidirectional soft alignment between image and encoded disease-label representations. This design reduces inappropriate repulsion between non-paired samples with related disease-label patterns and encourages the image encoder to preserve disease-aware inter-sample relations. During prediction, Neighborhood-Prior Decision Calibration (NPDC) converts continuous disease probabilities into binary labels by estimating sample-label-specific thresholds and refining them with local neighborhood disease priors retrieved from an embedding-based training memory bank. Experiments using CheXpert development data with held-out CheXlocalize testing, together with independent experiments on NIH Chest X-rays, demonstrate that MRCA improves disease-aware representation learning and score-level discrimination, while NPDC provides a balanced threshold-dependent prediction profile by jointly considering sample-level consistency and label-wise decision quality. Under the CheXpert evaluation protocol, the proposed framework achieves the best Exact Match, Macro-AUROC, Macro-AUPRC, and Macro-F1 among the main comparison strategies. On NIH Chest X-rays, it obtains the highest Macro-AUROC, Macro-AUPRC, and Macro-F1, demonstrating consistent effectiveness on an additional dataset with imbalanced multi-label distributions.

KEYWORDS: Disease-label-guided representation learning; multi-label classification; relation-aware contrastive alignment; decision calibration

1 Introduction

Chest X-ray imaging is one of the most widely used diagnostic tools in clinical practice and plays an important role in computer-aided diagnosis [1,2]. Beyond conventional image-level disease classification, a clinically meaningful objective is to infer structured disease fields from radiographs, thereby supporting structured reporting, downstream clinical analysis, and decision support. In this work, we study chest X-ray multi-label prediction with structured disease-label-guided contrastive representation learning. During pre-training, report-derived binary disease-label fields are encoded as an auxiliary representation and aligned with the corresponding image representation. After pre-training, the auxiliary tabular branch is removed, and inference relies only on chest X-ray images.

Conventional image-only supervised learning uses disease annotations primarily as prediction targets and may underutilize the cross-sample relational structure contained in multi-label disease patterns. Report-derived disease-label fields provide a compact structured description of these annotations. Encoding them as auxiliary representations enables disease-label co-occurrence and inter-sample relationships to provide additional supervision for image representation learning.

Recent studies on tabular- or metadata-guided visual representation learning have shown that structured non-image information can provide effective supervision for image encoders. For example, CHARMS transfers expert-derived tabular knowledge to image classifiers by aligning image channels with tabular attributes, while metadata-enhanced contrastive learning incorporates patient metadata into medical image pre-training to improve downstream image-level tasks [3,4]. These studies suggest that structured clinical information can guide image encoders toward more clinically meaningful and discriminative representations. Thus, a disease-label-guided representation learning framework should exploit these structured fields as auxiliary relational supervision and establish a suitable alignment mechanism to guide the image representation space.

A widely adopted strategy for achieving cross-modal semantic alignment is to construct contrastive objectives between heterogeneous modalities. A representative example is CLIP [5], which aligns paired image-text representations while contrasting them against other samples in the batch. This instance-wise contrastive paradigm has inspired a broad line of medical vision-language pre-training methods. Early studies mainly focused on learning global alignment between radiographs and their corresponding reports. For example, ConVIRT [6] contrasts paired medical images and reports to improve visual representation learning. Subsequent methods further enhance cross-modal alignment by introducing more fine-grained semantic interactions. GLoRIA [7] models global and local correspondences between image regions and report words, while BioViL [8] leverages radiology report semantics to improve biomedical vision-language representations. Beyond strictly paired image-report learning, MedCLIP [9] incorporates medical semantic matching to enable contrastive learning from unpaired medical image-text data. Together, these studies show that auxiliary clinical semantics can be transferred to medical image encoders through cross-modal pre-training, leading to more informative visual representations for downstream tasks. Beyond image-text learning, recent studies have investigated image-tabular multimodal representation learning, where structured clinical variables provide complementary information to visual features. MMCL [10] extends contrastive learning to image-tabular data by combining image and tabular augmentations, enabling tabular information to guide the training of image encoders for downstream unimodal prediction. TIP [11] further

develops tabular–image pre-training for incomplete heterogeneous tabular data by introducing image–tabular contrastive learning, image–tabular matching, and masked tabular reconstruction. More recently, PMRL [12] revisits multimodal alignment from a more principled perspective and encourages simultaneous alignment across modalities without relying on a fixed anchor modality.

Despite these advances, most existing image–tabular alignment objectives are still mainly built upon instance-level correspondence. In conventional CLIP-style contrastive learning, the paired image and auxiliary modality record from the same subject are treated as the positive pair, while other samples are generally regarded as negatives or selected as hard negatives. This formulation is reasonable when non-paired samples are semantically distinct, but it becomes problematic in medical multi-label scenarios, where patients may share partially overlapping disease labels or similar clinical phenotypes [13–15]. As illustrated in Fig. 1, this binary positive–negative assumption may incorrectly treat clinically related non-paired patients as negatives and push their representations apart, thereby introducing false-negative contrastive supervision and weakening disease-aware cross-modal representation learning.

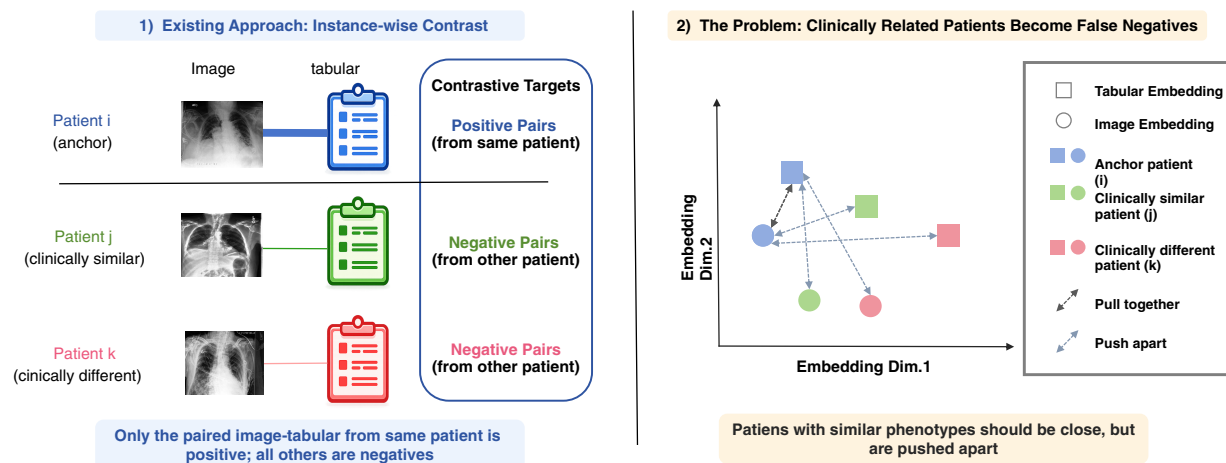


Figure 1: Illustration of false-negative supervision in conventional instance-wise image–tabular contrastive learning. Only the same-patient image–tabular pair is treated as positive, whereas clinically related non-paired patients may be incorrectly treated as negatives and pushed apart in the representation space.

Several contrastive learning variants have attempted to mitigate false-negative supervision by redefining positive samples or correcting the influence of false negatives. Supervised contrastive learning [16] pulls samples from the same class closer in the embedding space, while debiased contrastive learning [17] corrects the bias introduced by sampling false negatives from unlabeled data. MedCLIP [9] also reduces false negatives in medical image–text learning by introducing semantic matching based on medical knowledge. However, these methods are not specifically designed for graded relation construction in medical multi-label learning. In this setting, patient similarity is inherently graded rather than binary: different patients may share partially overlapping disease findings, comorbidity patterns, or similar clinical phenotypes, even when they are not identical paired samples. Therefore, cross-modal alignment should not only preserve the reliable same-patient image–tabular correspondence, but also account for disease-label overlap and phenotype-level similarity among different patients. Existing contrastive formulations still provide limited ability to explicitly encode such multi-relational clinical similarity as a soft cross-modal alignment target.

Beyond representation learning, chest X-ray multi-label prediction also faces a decision-level challenge during score-to-label conversion. After fine-tuning, the model produces continuous probabilities for each disease label, whereas clinical decision making requires binary disease labels. A common practice is to

apply a fixed threshold, such as 0.5, to all labels and patients. However, this strategy implicitly assumes a universal decision boundary across heterogeneous disease labels and patient cases, which is often unrealistic for medical multi-label data. Disease labels are typically imbalanced in chest X-ray multi-label classification, and different labels may have substantially different prevalence rates and decision thresholds [18–21]. At the patient level, disease manifestations and co-occurrence patterns may also vary across samples, indicating that the same prediction score does not necessarily imply the same clinical decision for different patients. Consequently, fixed-threshold decision making may introduce decision-boundary mismatch and produce unreliable binary predictions, especially for rare diseases or patient cases whose disease patterns deviate from the global training distribution. These characteristics highlight the need for a decision calibration strategy that can account for both label-level imbalance and patient-level heterogeneity.

Existing calibration and imbalance-aware methods partially address score reliability or label-distribution bias, but they do not explicitly model sample-label-specific decision boundaries for multi-label prediction. Temperature scaling [22] calibrates confidence scores through post-hoc rescaling, but it mainly improves probability calibration rather than determining label-wise decision thresholds. Label smoothing [23] replaces hard labels with softened targets and can improve generalization and calibration, yet it acts as a global training regularizer and does not adapt decision boundaries to different labels or patient cases. For imbalanced classification, logit adjustment [24] incorporates label priors into classifier logits to reduce long-tailed distribution bias, while distribution-balanced loss [25] further considers label co-occurrence and negative-label dominance in multi-label learning. However, these methods mainly intervene at the probability, target, logit, or loss level. They provide limited mechanisms for adapting the final score-to-label conversion to both label-level imbalance and patient-level disease heterogeneity. Therefore, a decision calibration strategy that estimates sample-label-specific thresholds is needed for reliable medical multi-label prediction.

The proposed framework is designed to mitigate false-negative cross-modal supervision during representation learning and boundary mismatch during score-to-label conversion. It consists of Multi-Relational Cross-Modal Alignment (MRCA) for disease-aware representation learning and Neighborhood-Prior Decision Calibration (NPDC) for sample-label-specific decision calibration. MRCA learns a disease-aware image representation space through alignment with encoded disease-label representations, while NPDC further exploits the resulting embedding neighborhood to calibrate sample-label-specific decision thresholds.

In the pre-training stage, we introduce MRCA, which replaces identity-only contrastive targets with relation-aware soft targets constructed from patient identity, disease-label similarity, and tabular phenotype similarity. By assigning non-zero alignment weights to clinically related non-paired patients, MRCA reduces false-negative contrastive supervision and encourages the image encoder to capture disease-aware clinical structures. In this way, report-derived disease-label fields provide auxiliary relational guidance for visual representation learning, while the final inference process remains image-only after the auxiliary tabular branch is removed.

In the prediction stage, the pre-trained image encoder is transferred to downstream image-only multi-label prediction. The model first produces continuous disease probabilities from chest X-ray images, and NPDC is then introduced for threshold-sensitive inference. Instead of relying on a universal threshold or only applying global logit correction, NPDC estimates sample-label-specific thresholds from prediction features, uncertainty cues, and global disease priors. These thresholds are further refined using local neighborhood disease priors retrieved from the embedding neighborhood in the training memory bank. When a query image is close to training samples with a higher prevalence of a certain disease, the corresponding threshold can be adaptively reduced to improve sensitivity; otherwise, it can be increased to suppress potential false positives. Through this representation-to-decision pipeline, the proposed framework uses structured

disease-label relations to guide image representation learning and further adapts score-to-label conversion to label imbalance and patient-level heterogeneity.

The main contributions of this work are summarized as follows:

- A chest X-ray multi-label prediction setting with structured disease-label-guided contrastive representation learning is formulated. Report-derived disease-label fields provide auxiliary relational supervision during pre-training, while inference relies only on chest X-ray images.
- To mitigate false-negative contrastive supervision, MRCA constructs relation-aware soft targets from paired identity, disease-label similarity, and tabular phenotype similarity.
- To alleviate decision-boundary mismatch during score-to-label conversion, Neighborhood-Prior Decision Calibration (NPDC) is introduced for threshold-sensitive multi-label inference. NPDC estimates sample-label-specific thresholds from prediction features, uncertainty cues, and global disease priors, and further refines them with local neighborhood disease priors retrieved from embedding neighborhoods.
- Extensive experiments on CheXpert and NIH Chest X-rays demonstrate the effectiveness of the proposed framework. The results show that MRCA improves disease-aware representation learning and score-level discrimination, while NPDC provides a balanced threshold-dependent prediction profile by jointly considering sample-level consistency and label-wise decision quality under imbalanced disease distributions.

2 Methodology

2.1 Problem Definition

Given a chest X-ray image x_i of sample i , the task is to predict a C -dimensional binary disease-label vector, where each element indicates whether a specific disease label is present. The image encoder and prediction head first estimate a disease probability vector $\hat{\mathbf{p}}_i$, as shown in Eq. (1):

$$\hat{\mathbf{p}}_i = [\hat{p}_i^1, \hat{p}_i^2, \dots, \hat{p}_i^C] = \text{Norm}(g(f_I(x_i))) \in [0, 1]^C, \quad (1)$$

where $f_I(\cdot)$ denotes the image encoder, $g(\cdot)$ denotes the prediction head, C is the number of disease labels, and $\text{Norm}(\cdot)$ denotes label-wise probability normalization. In this work, $\text{Norm}(\cdot)$ is implemented by applying softmax normalization to the two-class logits of each disease label. The element $\hat{p}_i^c$ represents the predicted positive-class probability of sample i for disease label c .

To obtain a binary prediction, each disease probability is compared with a decision threshold. For disease label c , the prediction of sample i is defined as:

$$\hat{y}_i^c = \mathbb{I}(\hat{p}_i^c \geq \theta_i^c), \quad (2)$$

where θ_i^c denotes the decision threshold for disease label c of sample i , and $\mathbb{I}(\cdot)$ is the indicator function. When $\hat{p}_i^c \geq \theta_i^c$, sample i is predicted to be positive for disease label c ; otherwise, it is predicted to be negative.

The resulting prediction vector is $\hat{\mathbf{y}}_i = [\hat{y}_i^1, \dots, \hat{y}_i^C] \in \{0, 1\}^C$, and the corresponding ground-truth vector is $\mathbf{y}_i = [y_i^1, \dots, y_i^C] \in \{0, 1\}^C$, where $y_i^c = 1$ indicates the presence of disease label c and $y_i^c = 0$ otherwise.

This study considers chest X-ray multi-label prediction with structured disease-label-guided representation learning. During pre-training, each sample is associated with a chest X-ray image and report-derived binary disease-label fields organized in a tabular form. An auxiliary tabular encoder transforms these fields into a structured disease-label representation that is used to characterize inter-sample relationships and guide image representation learning. During inference, the auxiliary tabular branch is removed, and the final model predicts the disease vector from the chest X-ray image alone. Under this formulation, the

central objective is to learn an image encoder that can exploit structured disease-label relations during pre-training and then support image-only multi-label prediction at deployment. Therefore, the learned image representation should preserve the original same-patient image–tabular correspondence while also reflecting clinically meaningful disease-label and phenotype-level relations among different patients. After the image encoder produces disease probabilities, the model further needs to convert these continuous scores into binary disease labels in a way that accounts for label imbalance and patient-level heterogeneity. Accordingly, the proposed methodology is organized into two stages. The first stage performs disease-aware cross-modal representation learning, where Multi-Relational Cross-Modal Alignment (MRCA) is constructed as the representation learning objective to align image and tabular disease-label representations using structured disease-label relations as soft supervision. The second stage performs image-only multi-label prediction with Neighborhood-Prior Decision Calibration (NPDC), where sample-label-specific thresholds are calibrated to obtain reliable binary disease predictions.

2.2 Overall Framework

The proposed framework provides a unified solution for chest X-ray multi-label prediction with structured disease-label-guided representation learning. As illustrated in [Fig. 2](#), it consists of a representation learning stage and a decision calibration stage. **Stage (1): Disease-Aware Cross-Modal Representation Learning.** This stage aims to transfer structured disease semantics from report-derived tabular disease-label fields into the image encoder during pre-training. Given paired chest X-ray images and tabular disease-label fields, the image encoder and tabular encoder first project the two representations into a shared representation space. To learn clinically meaningful visual representations, Multi-Relational Cross-Modal Alignment (MRCA) is introduced as the core pre-training objective. Instead of relying only on same-patient image–tabular correspondence, MRCA constructs relation-aware soft alignment targets by jointly modeling patient identity, disease-label similarity, and tabular phenotype similarity. By assigning non-zero alignment weights to clinically related non-paired samples, MRCA reduces false-negative contrastive supervision and encourages the learned image representation space to preserve disease-aware patient relationships ([Section 2.3](#)). **Stage (2): Multi-Label Prediction with Neighborhood-Prior Decision Calibration.** After pre-training, the auxiliary tabular branch is discarded, and the pre-trained image encoder is transferred to downstream image-only multi-label prediction. The prediction head first produces continuous disease probabilities from chest X-ray images. Neighborhood-Prior Decision Calibration (NPDC) then converts these probabilities into binary disease labels by estimating sample-label-specific thresholds from prediction features, global disease priors, and local neighborhood disease priors retrieved from the embedding space ([Section 2.4](#)). Overall, the proposed framework first learns disease-aware image representations through auxiliary image–tabular alignment and then calibrates score-to-label conversion under label imbalance and patient-level heterogeneity.

2.3 Disease-Aware Cross-Modal Representation Learning

During pre-training, paired chest X-ray images and tabular clinical disease fields are used to learn a disease-aware image–tabular representation space. The goal of this stage is not limited to aligning each image with its paired tabular record. More importantly, structured disease semantics from tabular clinical fields are transferred into the image encoder, so that the learned image representations can better reflect disease-level and phenotype-level clinical structures. This representation learning process is particularly important for medical multi-label data, where different patients may exhibit similar disease combinations, comorbidity patterns, or clinical phenotypes, even though they are not identical paired samples. However, conventional

instance-level contrastive learning usually adopts a hard one-to-one matching target: only the paired image–tabular record is regarded as positive, while all other samples in the mini-batch are treated as negatives. Such a strict matching assumption may introduce false-negative contrastive supervision by pushing clinically related patients apart in the representation space.

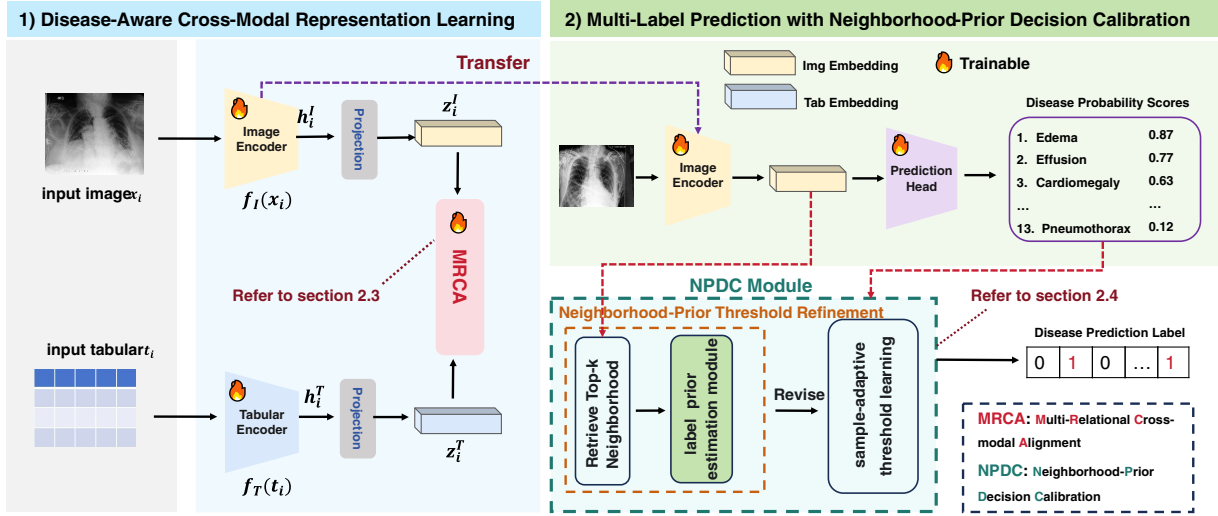


Figure 2: General architecture of the proposed framework, illustrating auxiliary image–disease-label alignment during pre-training and subsequent image-only multi-label prediction. The framework incorporates multi-relational cross-modal alignment to learn disease-aware image representations and neighborhood-prior decision calibration to produce reliable calibrated disease labels.

To achieve disease-aware cross-modal representation learning, we construct the Multi-Relational Cross-Modal Alignment (MRCA) objective, as illustrated in Fig. 3. MRCA provides the representation-level supervision for learning a clinically structured image–tabular embedding space. Instead of using a binary positive–negative target, MRCA first constructs a clinical multi-relation target by jointly modeling paired identity relation, disease-label similarity, and tabular phenotype similarity. The paired image–tabular record is retained as the strongest positive relation, while clinically related non-paired samples are assigned non-zero alignment weights according to their disease-label and phenotype similarities. The clinical multi-relation target is then converted into a soft target distribution and used to supervise bidirectional soft cross-modal contrastive learning. In this way, the MRCA objective guides representation learning with graded clinical relations rather than strict one-to-one correspondence, thereby reducing inappropriate repulsion among semantically related patients and promoting disease-aware image representation learning.

Given a mini-batch of paired image–tabular samples $\mathcal{B} = \{(x_i, \mathbf{t}_i, \mathbf{y}_i)\}_{i=1}^B$, where x_i denotes the chest X-ray image of sample i , $\mathbf{t}_i$ denotes its structured tabular disease-field input, and $\mathbf{y}_i \in \{0, 1\}^C$ denotes the corresponding multi-label disease annotation with C disease labels, the MRCA objective is constructed through two consecutive components. First, Multi-Relational Disease Target Construction estimates the clinical affinity between samples and transforms it into a sparse soft target distribution q_{ij} , which serves as the representation-level supervision signal. Second, Bidirectional Soft Cross-Modal Contrastive Learning uses q_{ij} to optimize image-to-tabular and tabular-to-image alignment in the shared representation space. Together, these two components define the MRCA objective for disease-aware cross-modal representation learning, enabling the learned representation space to preserve both same-patient correspondence and clinically meaningful patient-neighborhood structures.

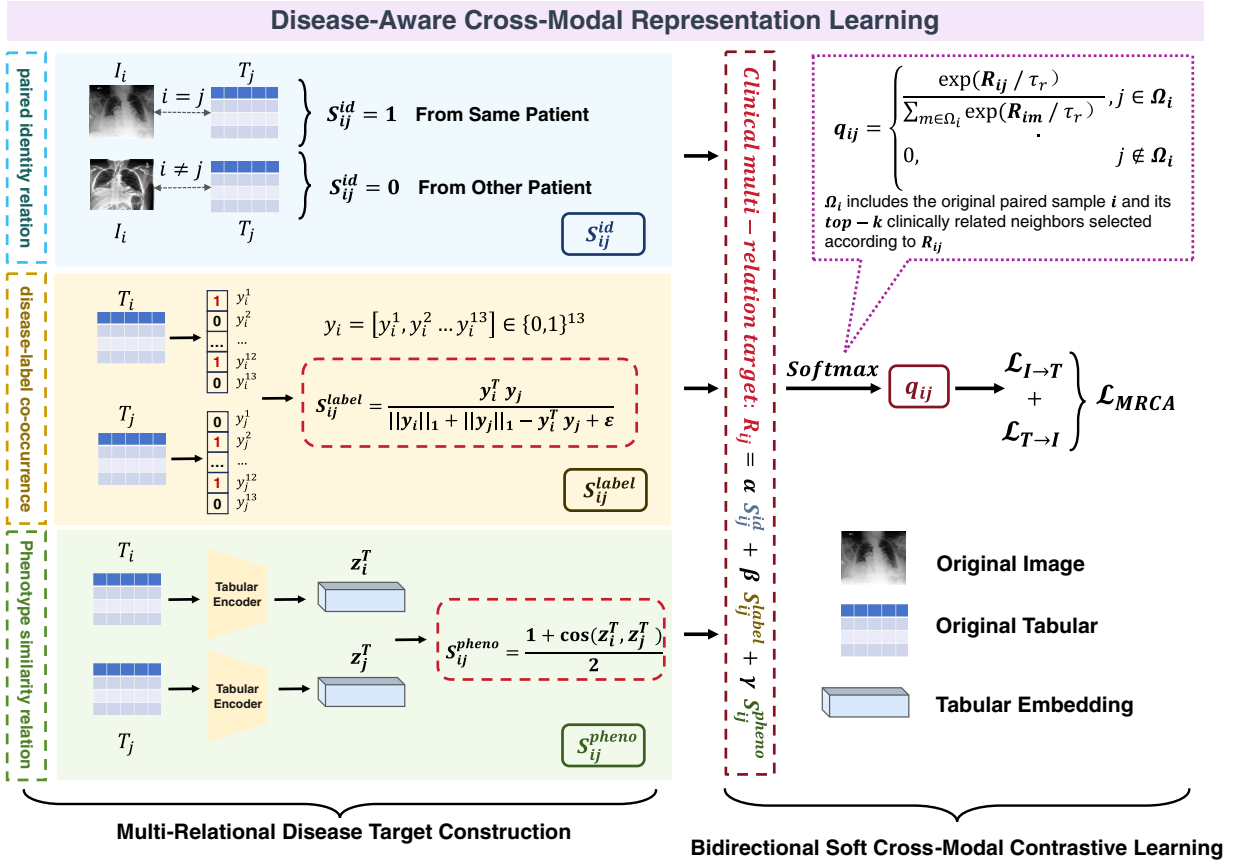


Figure 3: Illustration of the proposed disease-aware cross-modal representation learning stage. Multi-Relational Cross-Modal Alignment (MRCA) targets are constructed from paired identity relation, disease-label similarity, and tabular phenotype similarity and are then normalized into soft target distributions for bidirectional soft cross-modal contrastive learning. This design reduces false-negative supervision among clinically related patients and encourages disease-aware image–tabular representation learning.

2.3.1 Multi-Relational Disease Target Construction

The first component of the MRCA objective constructs the clinical supervision target for disease-aware representation learning. For each anchor sample i , its clinical relationship with other samples j in the mini-batch is estimated from three complementary perspectives: paired identity relation, disease-label similarity, and tabular phenotype similarity. These relations jointly describe whether two samples correspond to the same patient, whether they share common disease fields, and whether they exhibit similar phenotype-level clinical patterns. By encoding these relations into the supervision target, MRCA provides structured clinical guidance for learning a disease-aware representation space.

For an anchor image sample i , the paired identity relation S_{ij}^{id} indicates whether the image sample i and the tabular sample j correspond to the same patient. It is used to preserve the original image–tabular pairing, as shown in Eq. (3):

$$S_{ij}^{id} = \mathbb{I}(i = j). \quad (3)$$

Here, $S_{ij}^{id} = 1$ means that the image sample i and the tabular sample j come from the same patient. This relation anchors the contrastive objective by retaining the original same-patient image–tabular pair as the most reliable cross-modal positive pair.

The disease-label relation models explicit disease-level semantic overlap between patients. Given the multi-label disease vectors $\mathbf{y}_i \in \{0,1\}^C$ and $\mathbf{y}_j \in \{0,1\}^C$, their Jaccard similarity S_{ij}^{label} is computed as shown in Eq. (4):

$$S_{ij}^{label} = \frac{\mathbf{y}_i^\top \mathbf{y}_j}{\|\mathbf{y}_i\|_1 + \|\mathbf{y}_j\|_1 - \mathbf{y}_i^\top \mathbf{y}_j + \varepsilon}. \quad (4)$$

Here, $\mathbf{y}_i^\top \mathbf{y}_j$ denotes the number of shared positive disease labels, and ε is a small constant for numerical stability. This relation reflects the degree to which two patients share common disease fields. In medical multi-label data, patients with overlapping findings, such as cardiomegaly and pleural effusion, may exhibit related clinical semantics and therefore should not be treated as completely unrelated negatives in the representation learning process.

In addition to explicit disease-label overlap, tabular phenotype similarity is incorporated to capture broader patient-level clinical resemblance. Let $\mathbf{z}_i^T$ denote the tabular phenotype representation of patient i . The phenotype similarity relation S_{ij}^{pheno} is defined using rescaled cosine similarity, as shown in Eq. (5):

$$S_{ij}^{pheno} = \frac{1 + \cos(\mathbf{z}_i^T, \mathbf{z}_j^T)}{2}. \quad (5)$$

The cosine similarity is rescaled from $[-1, 1]$ to $[0, 1]$ to make it compatible with the other relation terms. Compared with explicit disease-label similarity, phenotype similarity provides a more general relational cue, allowing clinically similar patients to be identified even when their binary disease-label vectors are not exactly the same.

The final clinical multi-relation target is obtained by combining the three relation terms, as shown in Eq. (6):

$$R_{ij} = \alpha S_{ij}^{id} + \beta S_{ij}^{label} + \gamma S_{ij}^{pheno}. \quad (6)$$

Here, α , β , and γ control the contributions of paired identity, disease-label similarity, and tabular phenotype similarity, respectively. In the final configuration, α , β , and γ are set to 1.0, 0.1, and 0.7, respectively. This setting preserves the same-patient image–tabular pairing while assigning stronger emphasis to phenotype-level similarity, which is suitable for medical multi-label data where clinically similar patients may not always share identical binary disease-label vectors.

Although R_{ij} captures cross-sample clinical structure beyond identity-based pairing, not all samples in a mini-batch are clinically informative for a given anchor. If all samples are treated as soft positives, weak or noisy associations may be introduced into representation learning. Therefore, for each anchor sample i , only its top- K clinically related neighbors according to R_{ij} are retained, together with the original paired sample:

$$\Omega_i = \mathcal{N}_K(i) \cup \{i\}, \quad (7)$$

where $\mathcal{N}_K(i)$ denotes the set of top- K clinically related neighbors of sample i selected according to R_{ij} , and Ω_i denotes the resulting sparse relational positive set.

The retained relation scores are then normalized to obtain a soft target distribution q_{ij} , as shown in Eq. (8):

$$q_{ij} = \begin{cases} \frac{\exp(R_{ij}/\tau_r)}{\sum_{m \in \Omega_i} \exp(R_{im}/\tau_r)}, & j \in \Omega_i, \\ 0, & j \notin \Omega_i, \end{cases} \quad (8)$$

where τ_r is the relation temperature that controls the sharpness of the target distribution.

The resulting distribution q_{ij} serves as the soft supervision target of MRCA. It specifies the desired alignment strength between the image representation of anchor sample i and the tabular representation of sample j . Compared with the identity-based one-positive contrastive target, where only the paired sample is treated as positive, q_{ij} assigns non-zero probability mass to clinically related samples selected from the clinical multi-relation target. Thus, Multi-Relational Disease Target Construction provides disease-aware representation-level supervision for the subsequent bidirectional soft cross-modal contrastive learning.

2.3.2 Bidirectional Soft Cross-Modal Contrastive Learning

The second component of the MRCA objective transforms the soft target distribution q_{ij} into a contrastive representation learning objective. Different from conventional instance-level contrastive learning, which uses a hard identity-based matching target, MRCA supervises the contrastive objective with the soft target distribution derived from the clinical multi-relation target. Therefore, clinically related non-paired samples can contribute to the representation learning process with different alignment strengths, rather than being uniformly treated as negatives.

The image-to-tabular and tabular-to-image logits are computed in the shared embedding space as follows:

$$\begin{aligned} \ell_{ij}^{I \rightarrow T} &= \frac{(\tilde{\mathbf{z}}_i^I)^\top \tilde{\mathbf{z}}_j^T}{\tau}, \\ \ell_{ij}^{T \rightarrow I} &= \frac{(\tilde{\mathbf{z}}_i^T)^\top \tilde{\mathbf{z}}_j^I}{\tau}, \end{aligned} \quad (9)$$

where τ is the contrastive temperature.

Using the soft target distribution q_{ij} as supervision, the image-to-tabular loss $\mathcal{L}_{I \rightarrow T}$ is defined as:

$$\mathcal{L}_{I \rightarrow T} = -\frac{1}{B} \sum_{i=1}^B \sum_{j=1}^B q_{ij} \log \frac{\exp(\ell_{ij}^{I \rightarrow T})}{\sum_{m=1}^B \exp(\ell_{im}^{I \rightarrow T})}. \quad (10)$$

Similarly, the tabular-to-image loss $\mathcal{L}_{T \rightarrow I}$ is defined as:

$$\mathcal{L}_{T \rightarrow I} = -\frac{1}{B} \sum_{i=1}^B \sum_{j=1}^B q_{ij} \log \frac{\exp(\ell_{ij}^{T \rightarrow I})}{\sum_{m=1}^B \exp(\ell_{im}^{T \rightarrow I})}. \quad (11)$$

The final MRCA objective is obtained by averaging the two directional losses:

$$\mathcal{L}_{MRCA} = \frac{1}{2} (\mathcal{L}_{I \rightarrow T} + \mathcal{L}_{T \rightarrow I}). \quad (12)$$

Eq. (12) defines the optimization form of the MRCA objective for disease-aware cross-modal representation learning. The image-to-tabular direction encourages each image representation to align not only with its paired tabular record, but also with clinically related tabular samples according to the soft

2.4.1 Multi-Label Classifier Fine-Tuning

Given a chest X-ray image x_i , the transferred image encoder extracts $\mathbf{z}_i^I = f_I(x_i)$, and the multi-label prediction head produces a two-class logit vector $\ell_i^c = [\ell_{i,0}^c, \ell_{i,1}^c]$ for each disease label c . The positive-class probability $\hat{p}_i^c$ is obtained by applying label-wise softmax normalization to these logits.

The transferred image encoder and prediction head are fine-tuned using the mean cross-entropy loss over all samples and disease labels as shown in Eq. (13):

$$\mathcal{L}_{cls} = -\frac{1}{BC} \sum_{i=1}^B \sum_{c=1}^C \log \frac{\exp(\ell_{i,y_i^c}^c)}{\exp(\ell_{i,0}^c) + \exp(\ell_{i,1}^c)}, \quad (13)$$

where B is the mini-batch size, and $y_i^c \in \{0, 1\}$ denotes the preprocessed binary target for disease label c of sample i . The term $\ell_{i,y_i^c}^c$ denotes the logit associated with the target class, corresponding to the negative-class logit $\ell_{i,0}^c$ when $y_i^c = 0$ and the positive-class logit $\ell_{i,1}^c$ when $y_i^c = 1$.

After classifier fine-tuning, the image encoder and prediction head are fixed. The fixed classifier is subsequently used to produce the image embedding $\mathbf{z}_i^I$, the disease-wise logits ℓ_i , and the predicted probability vector $\hat{\mathbf{p}}_i$ for the threshold learning and neighborhood-prior calibration stages.

2.4.2 Sample-Adaptive Threshold Learning

A direct fixed-threshold strategy, such as $\theta = 0.5$, is often suboptimal for imbalanced medical multi-label prediction. Different disease labels may have different positive rates, and different patients may require different decision boundaries. Therefore, a sample-adaptive threshold module is introduced to estimate a disease-specific threshold for each validation or test sample.

For a validation sample i , the fixed image encoder and prediction head are used to obtain its image embedding $\mathbf{e}_i$, disease-wise logits ℓ_i^c , and predicted positive probability $\hat{p}_i^c$. These prediction features are combined with the global disease prior and uncertainty-related features to estimate an adaptive threshold:

$$\theta_{i,c}^{ada} = \theta_{\min} + (\theta_{\max} - \theta_{\min}) \sigma \left(h_{\phi}(\mathbf{e}_i, \ell_i^c, \hat{p}_i^c, \pi_c^{global}, u_i^c) \right), \quad (14)$$

where $h_{\phi}(\cdot)$ denotes the threshold estimation network. The constants $\theta_{\min}$ and $\theta_{\max}$ constrain the threshold to a valid interval. The term π_c^{global} denotes the global prior of disease label c , and u_i^c denotes the uncertainty-related feature vector.

The global disease prior π_c^{global} is computed from the training set as shown in Eq. (15):

$$\pi_c^{global} = \frac{1}{N_{tr}} \sum_{j=1}^{N_{tr}} y_j^c, \quad (15)$$

where N_{tr} is the number of training samples. Thus, π_c^{global} represents the positive ratio of disease label c in the training set.

The uncertainty-related feature vector is defined as:

$$u_i^c = [H(\hat{p}_i^c), |\hat{p}_i^c - 0.5|], \quad (16)$$

where the predictive entropy $H(\hat{p}_i^c)$ is computed as shown in Eq. (17):

$$H(\hat{p}_i^c) = -\hat{p}_i^c \log(\hat{p}_i^c + \varepsilon) - (1 - \hat{p}_i^c) \log(1 - \hat{p}_i^c + \varepsilon). \quad (17)$$

Here, $H(\hat{p}_i^c)$ measures prediction uncertainty, while $|\hat{p}_i^c - 0.5|$ measures the confidence margin from the default decision boundary.

Since hard thresholding is non-differentiable, a smooth approximation is used during threshold learning:

$$\tilde{y}_i^c = \sigma \left(\frac{\hat{p}_i^c - \theta_{i,c}^{ada}}{\tau_s} \right), \quad (18)$$

where τ_s controls the smoothness of the thresholding function. The soft output $\tilde{y}_i^c$ provides a differentiable approximation to the binary decision induced by $\theta_{i,c}^{ada}$.

The threshold estimation network is optimized using a differentiable macro-F1 surrogate $\mathcal{L}_{F1}$ computed from the soft outputs $\tilde{y}_i^c$. A binary cross-entropy term $\mathcal{L}_{BCE}$ is included to provide stable label-wise supervision. The resulting objective is:

$$\mathcal{L}_{ada} = \mathcal{L}_{F1} + \eta_{bce} \mathcal{L}_{BCE}, \quad (19)$$

where η_{bce} controls the contribution of the BCE term.

2.4.3 Neighborhood-Prior Threshold Refinement

Although the sample-adaptive threshold module estimates sample-specific thresholds from the prediction features of the current sample, it does not explicitly exploit neighborhood-level disease evidence. To address this limitation, a neighborhood-prior decision calibration strategy is introduced. As shown in Fig. 4, a training memory bank is first constructed after classifier fine-tuning, and local neighborhood disease priors are then estimated by retrieving similar training samples in the image embedding space. The resulting local priors are further compared with the global training-set prior to refine the adaptive decision threshold.

After classifier fine-tuning, the fixed image encoder extracts embeddings for all training samples. Each embedding $\mathbf{z}_j^I$ is stored with its ground-truth disease vector $\mathbf{y}_j$ to form the training memory bank $\mathcal{M}_{tr} = \{(\mathbf{z}_j^I, \mathbf{y}_j)\}_{j=1}^{N_{tr}}$. The memory bank is used only as the reference set for neighbor retrieval; validation and test labels are never included.

For a query sample i , its image embedding $\mathbf{e}_i$ is first obtained using the fixed image encoder. The top- K_n nearest training neighbors are then retrieved according to cosine similarity in the image embedding space:

$$\mathcal{N}_{K_n}(i) = \text{TopK}_{j \in \mathcal{D}_{tr}} \cos(\mathbf{z}_i^I, \mathbf{z}_j^I), \quad (20)$$

where $\mathcal{D}_{tr}$ denotes the training set, and $\mathcal{N}_{K_n}(i)$ denotes the set of retrieved training neighbors for query sample i .

To account for different neighbor relevance, the contribution of each retrieved neighbor is determined by a softmax over cosine similarities:

$$w_{ij} = \frac{\exp(\cos(\mathbf{z}_i^I, \mathbf{z}_j^I)/\tau_n)}{\sum_{m \in \mathcal{N}_{K_n}(i)} \exp(\cos(\mathbf{z}_i^I, \mathbf{z}_m^I)/\tau_n)}, \quad (21)$$

where w_{ij} denotes the normalized contribution of neighbor j to query sample i , and τ_n controls the sharpness of the neighbor-weight distribution.

For disease label c , the local disease prior of query sample i is computed as the weighted average of the labels of its retrieved neighbors:

$$\pi_{i,c}^{local} = \sum_{j \in \mathcal{N}_{K_n}(i)} w_{ij} y_j^c. \quad (22)$$

Here, y_j^c denotes the ground-truth label of training neighbor j for disease label c . Therefore, $\pi_{i,c}^{local}$ estimates the disease tendency of query sample i based on its local embedding neighborhood.

The local disease prior is then compared with the global training-set prior:

$$\Delta\pi_{i,c} = \pi_{i,c}^{local} - \pi_c^{global}, \quad (23)$$

where π_c^{global} denotes the positive ratio of disease label c in the training set. A positive $\Delta\pi_{i,c}$ indicates that disease label c is more prevalent in the local neighborhood of sample i than in the overall training set, whereas a negative value indicates a lower local disease tendency.

The neighborhood-prior calibrated threshold $\theta_{i,c}^{NPDC}$ is obtained by refining the sample-adaptive threshold with the prior discrepancy:

$$\theta_{i,c}^{NPDC} = \Pi_{[\theta_{\min}, \theta_{\max}]}(\theta_{i,c}^{ada} - \lambda_n \Delta\pi_{i,c}), \quad (24)$$

where $\Pi_{[\theta_{\min}, \theta_{\max}]}(\cdot)$ denotes the projection onto the interval $[\theta_{\min}, \theta_{\max}]$, and λ_n controls the strength of neighborhood-prior calibration. When $\Delta\pi_{i,c} > 0$, stronger neighborhood evidence is provided for disease label c , and the threshold is reduced to improve sensitivity. When $\Delta\pi_{i,c} < 0$, the threshold is increased to suppress potential false positives.

Finally, the binary prediction follows the decision rule in Eq. (2), with θ_i^c instantiated as the calibrated threshold $\theta_{i,c}^{NPDC}$.

2.5 Training and Inference Protocol

The proposed framework follows three sequential optimization stages, followed by memory-bank construction and image-only inference. First, MRCA jointly optimizes the image encoder and auxiliary tabular encoder using the relational structure derived from paired identity, disease-label similarity, and tabular phenotype similarity. Second, the auxiliary tabular branch is removed, and the image encoder is transferred to image-only classifier fine-tuning. Third, with the classifier fixed, the threshold network is learned on the validation set to estimate sample- and label-specific decision thresholds. After optimization, a training memory bank is constructed to provide local disease priors for NPDC during inference. The complete procedure is summarized in Algorithm 1.

These stages separate representation learning, classifier optimization, and decision calibration. The auxiliary tabular encoder is used only during MRCA pre-training and is removed before classifier fine-tuning.

Algorithm 1: Sequential optimization and image-only inference procedure of the proposed framework

Require: Paired training set $\mathcal{D}_{tr} = \{(x_i, \mathbf{t}_i, \mathbf{y}_i)\}$, validation set $\mathcal{D}_{val}$, and query images

Require: Relation weights α , β , and γ ; MRCA neighborhood size K ; NPDC parameters K_n , τ_n , and λ_n

Ensure: Image-only binary multi-label predictions

Stage 1: MRCA pre-training

1: **for** each mini-batch $\mathcal{B} \subset \mathcal{D}_{tr}$ **do**

2: Encode the chest X-rays and structured disease-label fields

(Continued)

Algorithm 1 (continued)

-
- 3: Compute S^{id} , S^{label} , and S^{pheno} and combine them into R
 - 4: Retain the top- K relational positives and normalize them into q
 - 5: Update the image and auxiliary tabular branches by minimizing $\mathcal{L}_{MRCA}$
 - 6: **end for**
 - 7: Remove the auxiliary tabular branch
 - Stage 2: Image-only classifier fine-tuning**
 - 8: **for** each mini-batch $\mathcal{B} \subset \mathcal{D}_{tr}$ **do**
 - 9: Update the image encoder and prediction head by minimizing $\mathcal{L}_{cls}$
 - 10: **end for**
 - 11: Fix the image encoder and prediction head
 - Stage 3: Sample-adaptive threshold learning**
 - 12: Compute the global disease prior π^{global} from $\mathcal{D}_{tr}$
 - 13: **for** each mini-batch $\mathcal{B} \subset \mathcal{D}_{val}$ **do**
 - 14: Extract fixed image embeddings, logits, probabilities, and uncertainty features
 - 15: Estimate $\theta_{i,c}^{ada}$ and update h_ϕ by minimizing $\mathcal{L}_{ada}$
 - 16: **end for**
 - Stage 4: Memory-bank construction and NPDC inference**
 - 17: Construct $\mathcal{M}_{tr}$ using the fixed embeddings and labels from $\mathcal{D}_{tr}$
 - 18: **for** each query image x_i **do**
 - 19: Compute $\mathbf{z}_i^I$, $\hat{\mathbf{p}}_i$, and $\theta_{i,c}^{ada}$
 - 20: Retrieve $\mathcal{N}_{K_n}(i)$ from $\mathcal{M}_{tr}$ and estimate $\pi_{i,c}^{local}$
 - 21: Compute $\Delta\pi_{i,c}$ and refine $\theta_{i,c}^{ada}$ into $\theta_{i,c}^{NPDC}$
 - 22: Output $\hat{y}_i^c = \mathbb{I}(\hat{p}_i^c \geq \theta_{i,c}^{NPDC})$
 - 23: **end for**
-

The threshold network is subsequently optimized on the validation set while the image encoder and prediction head remain fixed. The resulting image embeddings and training labels are then stored in the memory bank, from which NPDC retrieves neighboring training samples and estimates a local disease prior. Thus, each query is processed using only its chest X-ray, while the training memory bank supports threshold refinement without requiring tabular information from the query sample.

The parameters that directly control MRCA relation construction and NPDC neighborhood-prior refinement are summarized in [Table 1](#).

3 Experiments

3.1 Datasets

We evaluated the proposed framework on three public chest radiograph datasets: CheXpert, CheXlocalize, and NIH Chest X-rays. CheXpert was used for MRCA pre-training, downstream fine-tuning, internal validation, model selection, and threshold calibration. CheXlocalize, which is derived from CheXpert, was used as the expert-annotated held-out test set for quantitative evaluation of the models trained on CheXpert. NIH Chest X-rays was independently divided for model training, validation, and testing to evaluate the framework on an additional dataset. The dataset statistics and their roles in the experimental pipeline are summarized in [Table 2](#).

Table 1: Core method-specific hyperparameter settings.

Parameter	Role	Value
α	Paired identity relation weight	1.0
β	Disease-label relation weight	0.1
γ	Tabular phenotype relation weight	0.7
K	MRCA soft-positive neighborhood size	10
K_n	NPDC retrieved neighborhood size	10
λ_n	NPDC refinement strength	0.05

Table 2: Statistics and usage of the chest radiograph datasets.

Dataset	Images	Patients	Annotation Type	Usage in this Study
CheXpert	224,316	65,240	Report-derived observations with uncertainty labels	Pre-training, fine-tuning, and validation
CheXlocalize	668	500	Image-level labels and expert localization annotations	Held-out quantitative testing
NIH Chest X-rays	112,120	30,805	Report-derived disease labels	Independent training, validation, and testing

3.1.1 CheXpert

CheXpert is a large-scale chest radiograph dataset released by Stanford University [26]. It contains 224,316 chest radiographs from 65,240 patients, together with associated radiology reports. The images include both frontal and lateral views. A rule-based labeler was applied to the reports to extract 14 observations, each annotated as positive, negative, uncertain, or unmentioned. These report-derived observations provide the structured disease-label fields associated with the chest X-ray images. We further preprocessed the original CheXpert label states before model training. Positive labels were mapped to 1, while negative, uncertain, and unmentioned labels were mapped to 0. This binarization step converted the original four-state report-derived annotations into binary disease indicators. The Support Devices label was then excluded from the downstream disease-label set because it indicates the presence of medical devices rather than a thoracic disease finding. The remaining 13 observation fields were organized into a multi-hot disease vector for each image. Each entry in this vector indicates the presence or absence of one disease observation after preprocessing.

In this way, each CheXpert sample consisted of a chest radiograph paired with its preprocessed multi-hot disease-label vector. During MRCA pre-training, these binary disease vectors were used to construct disease-label relations between samples and to guide cross-modal representation learning. During downstream training and evaluation, the same vectors served as image-level multi-label prediction targets. No pixel-level lesion annotations were used for MRCA pre-training, downstream fine-tuning, model selection, or threshold calibration.

In our experiments, CheXpert was used for both MRCA pre-training and downstream fine-tuning. The CheXpert training split was used to optimize the model, while the CheXpert validation split was used

for internal validation, model selection, and threshold calibration. No CheXlocalize test images or labels were used during these stages. Following our experimental setting, demographic and acquisition-related attributes, including sex, age, frontal/lateral view, and AP/PA view, were excluded from the prediction fields. The final CheXpert label space contained 13 observation fields: No Finding, Enlarged Cardiomedastinum, Cardiomegaly, Lung Opacity, Lung Lesion, Edema, Consolidation, Atelectasis, Pneumothorax, Pleural Effusion, Pleural Other, Fracture, and Pneumonia. These retained label fields exhibit substantial class imbalance, with the proportion of samples annotated as positive for each label ranging from 1.57% to 47.26%. The corresponding positive sample counts and proportions are reported in Supplementary Table S1.

3.1.2 CheXlocalize

CheXlocalize is an expert-annotated chest X-ray localization benchmark built upon the CheXpert dataset [27]. It provides radiologist annotations for localizable chest radiographic findings. In this study, we used its test split, which contains 668 chest radiographs from 500 patients, as the held-out test set for quantitative evaluation of the models trained on CheXpert.

No CheXlocalize test images or labels were used before final evaluation, including during pre-training, fine-tuning, model selection, or threshold calibration. This separation prevents test-set leakage and ensures a strictly held-out evaluation. For consistency with CheXpert-based training, we retained the same disease-label space whenever the corresponding labels were available and followed the same binary label definition used for CheXpert. The evaluated label fields in this held-out set are similarly imbalanced, with the proportion of positive samples ranging from 0.90% to 46.41%; the complete distribution is provided in Supplementary Table S1.

3.1.3 NIH Chest X-Rays

The NIH Chest X-rays dataset, also known as ChestX-ray14, was released by the National Institutes of Health Clinical Center [1]. It contains 112,120 frontal-view chest radiographs from 30,805 patients. The image-level disease labels were automatically extracted from radiology reports using natural language processing. Since each image may be associated with multiple findings, this dataset provides an appropriate additional benchmark for multi-label chest disease classification.

NIH Chest X-rays includes 14 thoracic disease categories: Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, Pneumothorax, Consolidation, Edema, Emphysema, Fibrosis, Pleural Thickening, and Hernia. Images without detected abnormalities are marked as *No Finding*. Unlike CheXpert, NIH Chest X-rays does not provide uncertain or unmentioned label states. Therefore, NIH labels were converted into binary multi-label vectors according to the presence or absence of each disease category, and images labeled as *No Finding* were treated as negative for all disease categories. The 14 disease categories show substantial imbalance, with the proportion of positive samples per category ranging from 0.20% to 17.74%, while 53.84% of the images are labeled as *No Finding*. Complete label-wise statistics are reported in Supplementary Table S2.

3.2 Experimental Setup

We compare the proposed MRCA with four representative contrastive pre-training baselines using the auxiliary tabular disease-label fields: original TIP [11], PMRL [12], ITC [5], and MMCL [10]. For each pre-training strategy, three decision schemes are evaluated independently: fixed thresholding with a threshold of 0.5, logit tuning, and the proposed Neighborhood-Prior Decision Calibration (NPDC). The fixed-threshold setting serves as the standard decision baseline, while logit tuning is included as an additional calibration

baseline. In contrast, NPDC keeps the classifier outputs fixed and refines sample- and label-specific decision thresholds by incorporating local neighborhood disease priors from the training memory bank.

We report Exact Match, Macro-AUROC, Macro-AUPRC, Macro-F1, and Micro-F1. All quantitative experiments were repeated using random seeds 2022, 2023, and 2024, and the results are reported as the mean $\pm$ sample standard deviation. Given N test samples and C disease labels, let $\mathbf{y}_i \in \{0, 1\}^C$ and $\hat{\mathbf{y}}_i \in \{0, 1\}^C$ denote the ground-truth and predicted binary multi-label vectors of sample i , respectively. Exact Match is computed as:

$$\text{Exact Match} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(\hat{\mathbf{y}}_i = \mathbf{y}_i), \quad (25)$$

where $\mathbb{I}(\cdot)$ is the indicator function. For each disease label c , AUROC_c and AUPRC_c are computed using the predicted probabilities $\{\hat{p}_i^c\}_{i=1}^N$ and the corresponding binary labels $\{y_i^c\}_{i=1}^N$. Macro-AUROC and Macro-AUPRC are calculated by averaging the label-wise AUROC and AUPRC values over all disease labels:

$$\text{Macro-AUROC} = \frac{1}{C} \sum_{c=1}^C \text{AUROC}_c, \quad (26)$$

$$\text{Macro-AUPRC} = \frac{1}{C} \sum_{c=1}^C \text{AUPRC}_c. \quad (27)$$

Since the three decision schemes mainly affect the conversion from probability scores to binary predictions, the threshold-dependent metrics, including Exact Match, Macro-F1, and Micro-F1, are used to evaluate their decision-level effectiveness. Macro-AUROC and Macro-AUPRC are ranking-based metrics and remain unchanged when the underlying score ranking is preserved. Therefore, for the same pre-training strategy, identical Macro-AUROC and Macro-AUPRC values are reported across fixed thresholding, logit tuning, and NPDC.

3.3 Overall Performance Comparison

Tables 3 and 4 report the main comparisons under the CheXpert-trained/CheXlocalize-tested protocol and on the independently trained and evaluated NIH Chest X-rays dataset, respectively. We compare the proposed MRCA with four representative pre-training baselines, including ITC, TIP, MMCL, and PMRL. For each pre-training strategy, we evaluate three decision schemes: fixed thresholding with a threshold of 0.5, logit tuning, and the proposed Neighborhood-Prior Decision Calibration (NPDC). NPDC keeps the classifier outputs unchanged and refines sample- and label-specific thresholds using local neighborhood disease priors from the training memory bank.

On the held-out CheXlocalize test split under the CheXpert evaluation protocol, MRCA achieves the best overall performance under fixed thresholding, obtaining the highest Exact Match, Macro-AUROC, Macro-AUPRC, Macro-F1, and Micro-F1 among all pre-training strategies. This demonstrates that multi-relational alignment between image and encoded disease-label representations improves both score-level discrimination and threshold-dependent multi-label prediction. After applying NPDC, MRCA further achieves the best Exact Match, Macro-AUROC, Macro-AUPRC, and Macro-F1. Although MMCL with NPDC obtains a slightly higher Micro-F1, MRCA with NPDC provides the most balanced performance across ranking-based and threshold-dependent metrics.

Table 3: Main comparison of disease-label-guided representation learning methods with different decision schemes under the CheXpert-trained and CheXlocalize-tested evaluation protocol. All models are evaluated on the expert-annotated CheXlocalize held-out test split. Results are reported as mean $\pm$ standard deviation across repeated runs. The best and second-best mean values for each metric are highlighted in bold and underlined, respectively. Macro-AUROC and Macro-AUPRC remain unchanged across decision schemes because the underlying prediction scores are fixed.

Method	Decision	Exact-Match	Macro-AUROC	Macro-AUPRC	Macro-F1	Micro-F1
ITC [5]	Fixed 0.5	0.0968 $\pm$ 0.0090	0.7695 $\pm$ 0.0079	0.4054 $\pm$ 0.0173	0.1950 $\pm$ 0.0190	0.4026 $\pm$ 0.0066
ITC	Logit tuning [24]	0.0423 $\pm$ 0.0047	0.7695 $\pm$ 0.0079	0.4054 $\pm$ 0.0173	0.3508 $\pm$ 0.0076	0.4587 $\pm$ 0.0132
ITC	Proposed NPDC	0.0733 $\pm$ 0.0207	0.7695 $\pm$ 0.0079	0.4054 $\pm$ 0.0173	0.3543 $\pm$ 0.0111	0.5479 $\pm$ 0.0076
TIP [11]	Fixed 0.5	0.0852 $\pm$ 0.0137	0.7411 $\pm$ 0.0035	0.3824 $\pm$ 0.0075	0.1426 $\pm$ 0.0103	0.3726 $\pm$ 0.0026
TIP	Logit tuning [24]	0.0103 $\pm$ 0.0067	0.7411 $\pm$ 0.0035	0.3824 $\pm$ 0.0075	0.3408 $\pm$ 0.0146	0.4187 $\pm$ 0.0155
TIP	Proposed NPDC	0.0893 $\pm$ 0.0164	0.7411 $\pm$ 0.0035	0.3824 $\pm$ 0.0075	0.3614 $\pm$ 0.0132	0.5187 $\pm$ 0.0091
MMCL [10]	Fixed 0.5	0.0752 $\pm$ 0.0086	<u>0.7812 $\pm$ 0.0085</u>	<u>0.4075 $\pm$ 0.0021</u>	0.1726 $\pm$ 0.0149	0.4226 $\pm$ 0.0037
MMCL	Logit tuning [24]	0.0373 $\pm$ 0.0018	<u>0.7812 $\pm$ 0.0085</u>	<u>0.4075 $\pm$ 0.0021</u>	0.3759 $\pm$ 0.0073	0.4496 $\pm$ 0.0155
MMCL	Proposed NPDC	<u>0.1207 $\pm$ 0.0113</u>	<u>0.7812 $\pm$ 0.0085</u>	<u>0.4075 $\pm$ 0.0021</u>	0.3801 $\pm$ 0.0042	0.5687 $\pm$ 0.0071
PMRL [12]	Fixed 0.5	0.0902 $\pm$ 0.0046	0.7667 $\pm$ 0.0137	0.3975 $\pm$ 0.0067	0.1713 $\pm$ 0.0121	0.3854 $\pm$ 0.0137
PMRL	Logit tuning [24]	0.0273 $\pm$ 0.0090	0.7667 $\pm$ 0.0137	0.3975 $\pm$ 0.0067	0.3601 $\pm$ 0.0063	0.4496 $\pm$ 0.0055
PMRL	Proposed NPDC	0.0897 $\pm$ 0.0231	0.7667 $\pm$ 0.0137	0.3975 $\pm$ 0.0067	0.3641 $\pm$ 0.0127	0.5413 $\pm$ 0.0084
MRCA (Ours)	Fixed 0.5	0.1192 $\pm$ 0.0097	0.8279 $\pm$ 0.0132	0.4867 $\pm$ 0.0046	0.2313 $\pm$ 0.0135	0.4699 $\pm$ 0.0047
MRCA (Ours)	Logit tuning [24]	0.0571 $\pm$ 0.0023	0.8279 $\pm$ 0.0132	0.4867 $\pm$ 0.0046	<u>0.3900 $\pm$ 0.0139</u>	0.4876 $\pm$ 0.0235
MRCA (Ours)	Proposed NPDC	0.1401 $\pm$ 0.0065	0.8279 $\pm$ 0.0132	0.4867 $\pm$ 0.0046	0.4147 $\pm$ 0.0014	<u>0.5537 $\pm$ 0.0107</u>

Table 4: Main comparison of disease-label-guided representation learning methods with different decision schemes on NIH Chest X-rays. Results are reported as mean $\pm$ standard deviation across repeated runs. The best and second-best mean values for each metric are highlighted in bold and underlined, respectively. Macro-AUROC and Macro-AUPRC remain unchanged across decision schemes because the within-label score rankings are preserved.

Method	Decision	Exact-Match	Macro-AUROC	Macro-AUPRC	Macro-F1	Micro-F1
ITC [5]	Fixed 0.5	0.5055 $\pm$ 0.0018	0.7765 $\pm$ 0.0129	0.1853 $\pm$ 0.0011	0.0695 $\pm$ 0.0043	0.1331 $\pm$ 0.0066
ITC	Logit tuning [24]	0.3736 $\pm$ 0.0223	0.7765 $\pm$ 0.0129	0.1853 $\pm$ 0.0011	0.2014 $\pm$ 0.0026	0.3062 $\pm$ 0.0092
ITC	Proposed NPDC	0.4322 $\pm$ 0.0117	0.7765 $\pm$ 0.0129	0.1853 $\pm$ 0.0011	0.2361 $\pm$ 0.0059	0.2900 $\pm$ 0.0084
TIP [11]	Fixed 0.5	<u>0.5031 $\pm$ 0.0043</u>	<u>0.7965 $\pm$ 0.0043</u>	0.1801 $\pm$ 0.0125	0.0701 $\pm$ 0.0067	0.1531 $\pm$ 0.0087
TIP	Logit tuning [24]	0.3521 $\pm$ 0.0131	<u>0.7965 $\pm$ 0.0043</u>	0.1801 $\pm$ 0.0125	0.1987 $\pm$ 0.0094	0.3021 $\pm$ 0.0110
TIP	Proposed NPDC	0.4221 $\pm$ 0.0098	<u>0.7965 $\pm$ 0.0043</u>	0.1801 $\pm$ 0.0125	0.2287 $\pm$ 0.0068	0.2957 $\pm$ 0.0217
MMCL [10]	Fixed 0.5	0.4931 $\pm$ 0.0163	0.7944 $\pm$ 0.0061	<u>0.1867 $\pm$ 0.0009</u>	0.0753 $\pm$ 0.0082	0.1521 $\pm$ 0.0033
MMCL	Logit tuning [24]	0.3721 $\pm$ 0.0048	0.7944 $\pm$ 0.0061	<u>0.1867 $\pm$ 0.0009</u>	<u>0.2497 $\pm$ 0.0035</u>	<u>0.3241 $\pm$ 0.0058</u>
MMCL	Proposed NPDC	0.4334 $\pm$ 0.0014	0.7944 $\pm$ 0.0061	<u>0.1867 $\pm$ 0.0009</u>	0.1921 $\pm$ 0.0112	0.2651 $\pm$ 0.0158
PMRL [12]	Fixed 0.5	0.4988 $\pm$ 0.0050	0.7721 $\pm$ 0.0134	0.1579 $\pm$ 0.0073	0.0553 $\pm$ 0.0006	0.1167 $\pm$ 0.0084
PMRL	Logit tuning [24]	0.3782 $\pm$ 0.0019	0.7721 $\pm$ 0.0134	0.1579 $\pm$ 0.0073	0.1831 $\pm$ 0.0041	0.3031 $\pm$ 0.0014
PMRL	Proposed NPDC	0.4307 $\pm$ 0.0087	0.7721 $\pm$ 0.0134	0.1579 $\pm$ 0.0073	0.2031 $\pm$ 0.0107	0.2731 $\pm$ 0.0052
MRCA (Ours)	Fixed 0.5	0.5011 $\pm$ 0.0071	0.8074 $\pm$ 0.0063	0.1899 $\pm$ 0.0062	0.0653 $\pm$ 0.0018	0.1467 $\pm$ 0.0025
MRCA (Ours)	Logit tuning [24]	0.3927 $\pm$ 0.0069	0.8074 $\pm$ 0.0063	0.1899 $\pm$ 0.0062	0.2031 $\pm$ 0.0047	0.3269 $\pm$ 0.0074
MRCA (Ours)	Proposed NPDC	0.4457 $\pm$ 0.0040	0.8074 $\pm$ 0.0063	0.1899 $\pm$ 0.0062	0.2531 $\pm$ 0.0007	0.3107 $\pm$ 0.0033

On the independently evaluated NIH Chest X-rays dataset, MRCA also shows consistent effectiveness. It achieves the highest Macro-AUROC and Macro-AUPRC among all pre-training strategies, indicating superior disease-wise ranking capability on this additional dataset. With NPDC, MRCA obtains the highest Macro-F1 and the best Micro-F1 among the NPDC-calibrated methods, suggesting that neighborhood-prior threshold calibration improves label-wise prediction quality under imbalanced multi-label distributions. The Exact Match of MRCA with NPDC is lower than that obtained under fixed thresholding on NIH. This is likely because Exact Match requires all labels of a sample to be predicted correctly and may favor

conservative predictions in highly imbalanced multi-label data. In contrast, the improvements in Macro-F1, Macro-AUROC, and Macro-AUPRC indicate stronger disease-wise recognition and score-ranking ability.

Macro-AUROC and Macro-AUPRC remain unchanged across different decision schemes within the same pre-training strategy because these metrics evaluate score-ranking quality and are independent of the final thresholding rule. Therefore, changes in Exact Match, Macro-F1, and Micro-F1 mainly reflect the effectiveness of the decision schemes, while Macro-AUROC and Macro-AUPRC reflect the representation quality learned by each pre-training method.

3.4 Qualitative Evaluation and Clinical Plausibility Analysis

Although the quantitative results demonstrate the overall effectiveness of MRCA and NPDC, case-level visualization is necessary to examine whether the model relies on clinically meaningful image evidence and how decision calibration affects individual predictions. We therefore conduct qualitative analysis from three complementary perspectives: representative prediction cases, disease-wise Grad-CAM visualization, and failure modes and clinical limitations.

3.4.1 Representative Prediction Cases

Fig. 5 presents representative NIH test cases comparing the fixed threshold of 0.5 with the proposed NPDC. In the selected cases, fixed thresholding either produces no positive prediction or identifies only part of the ground-truth label set. For example, it detects only Effusion in Sample S2713 while missing the co-occurring Atelectasis and Consolidation, and it fails to identify any positive finding in the remaining examples. In contrast, NPDC recovers the complete annotated label set in these cases, including both single-label findings, such as Cardiomegaly and Nodule, and multi-label combinations involving Atelectasis, Effusion, Infiltration, Consolidation, Edema, Pneumonia, and Emphysema. These cases suggest that sample-adaptive threshold estimation and neighborhood-prior refinement can help recover positive findings that are suppressed by the fixed threshold of 0.5. This effect is particularly relevant to multi-label cases, where weaker co-occurring findings may receive lower prediction scores than the dominant abnormality and are therefore more likely to be missed under a universal decision threshold. Nevertheless, accurately recognizing all co-occurring findings remains challenging because different thoracic abnormalities may exhibit subtle or overlapping radiographic patterns.

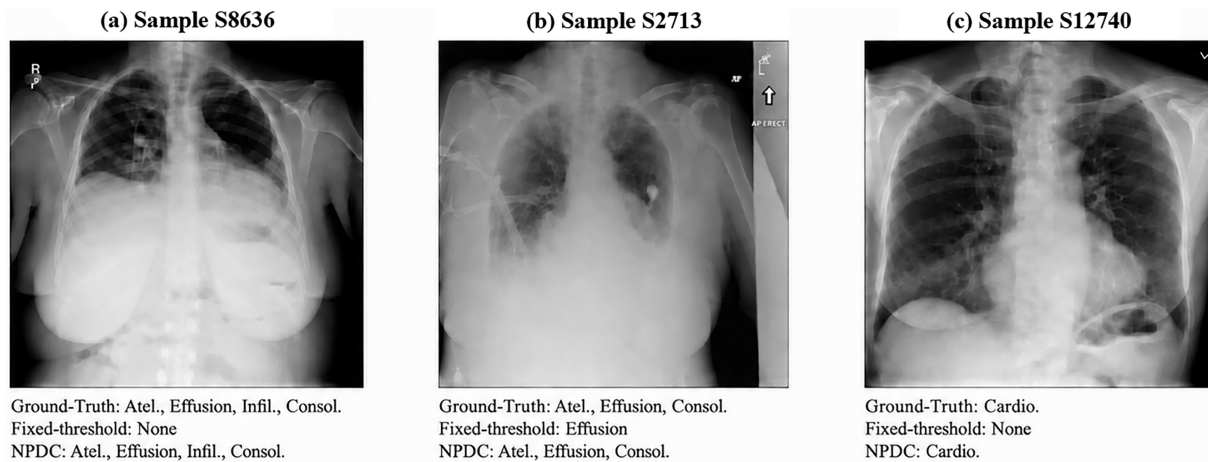


Figure 5: (Continued)

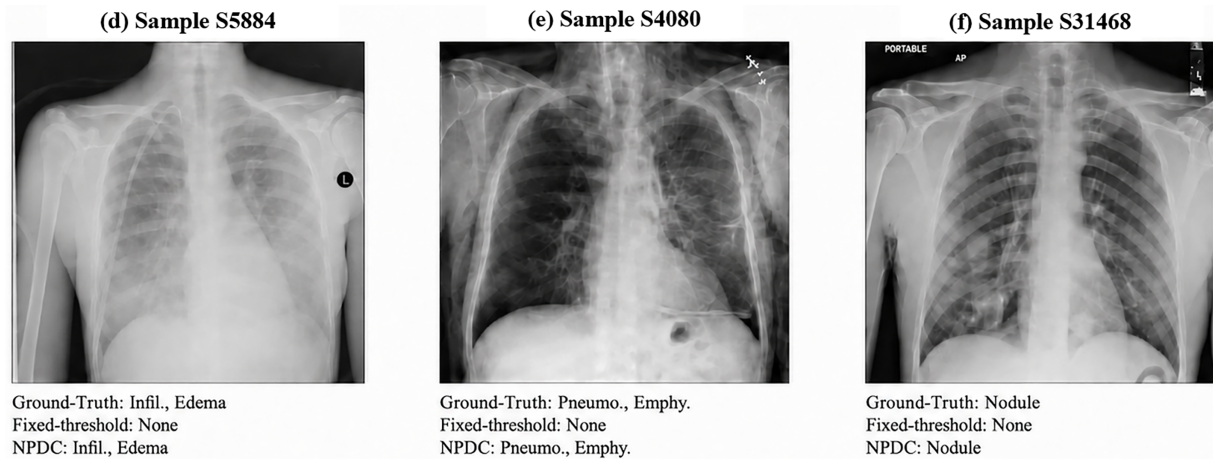


Figure 5: Representative prediction cases of MRCA with Neighborhood-Prior Decision Calibration (NPDC). Each panel shows the original chest X-ray, the ground-truth disease labels, the fixed-threshold prediction, and the NPDC prediction.

3.4.2 Disease-Wise Grad-CAM Visualization for Evaluation of Proposed Method

To further assess whether the predictions of the proposed method are supported by clinically plausible visual evidence, Fig. 6 presents disease-wise Grad-CAM visualizations obtained from the independently trained NIH model on representative NIH test samples. The visualized disease categories follow the NIH Chest X-rays label space. For each selected disease category, we show true-positive examples with the original chest X-ray and the corresponding Grad-CAM overlay. The attention responses for Pleural Effusion are mainly concentrated around the lower lung fields and pleural basal regions, while those for Cardiomegaly are primarily located around the enlarged cardiac silhouette. For Infiltration, Atelectasis, and Edema, the highlighted regions are mainly distributed within lung-field opacity areas. These disease-specific activation patterns indicate that, after multi-relational cross-modal alignment, the image encoder tends to focus on anatomically and pathologically relevant regions rather than arbitrary background areas. This observation supports the clinical plausibility of MRCA, suggesting that the transferred tabular disease semantics contribute to more disease-aware visual representation learning. It should be noted that these Grad-CAM results are used as qualitative evidence of clinical plausibility rather than as a quantitative localization evaluation, because the proposed framework is trained and evaluated under image-level supervision without using pixel-level lesion annotations.

3.4.3 Failure Modes and Clinical Limitations

Fig. 7 illustrates challenging examples in which the calibrated decision output remains inconsistent with the ground-truth labels. These cases are not intended as isolated prediction errors, but are used to characterize the residual limitations of image-only multi-label chest X-ray prediction under weak image-level supervision.

The failure cases reveal several representative error patterns. First, image-level annotations provide only disease presence or absence and do not indicate lesion locations. They may also be incomplete or affected by reporting uncertainty. As a result, images annotated as “No Finding” may still contain visually ambiguous basal opacities or pleural blunting-like patterns, which can lead to residual false-positive predictions, as shown in Fig. 7a. Second, different thoracic diseases may share overlapping radiographic manifestations. Opacity-related abnormalities, pleural findings, and focal lesions can present with subtle or partially overlapping visual cues, making it difficult to assign a unique disease label from the image alone. Third, decision calibration involves a trade-off between reducing noisy false positives and preserving weak

positive evidence. As shown in Fig. 7b and c, localized or subtle findings such as Mass and Pneumothorax may be suppressed when the corresponding visual evidence is weak after calibration.

These observations indicate that the remaining errors are not solely attributable to the decision calibration module. Instead, they reflect the broader difficulty of multi-label chest X-ray prediction with noisy, incomplete, and weakly localized image-level labels. In addition to these image-level limitations, the fixed binary-label formulation adopted in this study introduces a further source of label uncertainty. This formulation provides complete disease-label vectors and a consistent setting for evaluating multi-relational representation learning and decision calibration; however, mapping uncertain and unmentioned observations to the negative class may introduce false-negative label noise and affect explicit disease-label similarity, encoder-induced tabular phenotype similarity, and global or neighborhood disease-prior estimation. The reported results should therefore be interpreted under this binary-label assumption. Together, these factors explain why robust recognition of subtle, localized, and ambiguously annotated findings remains challenging even when the proposed method improves overall decision quality.

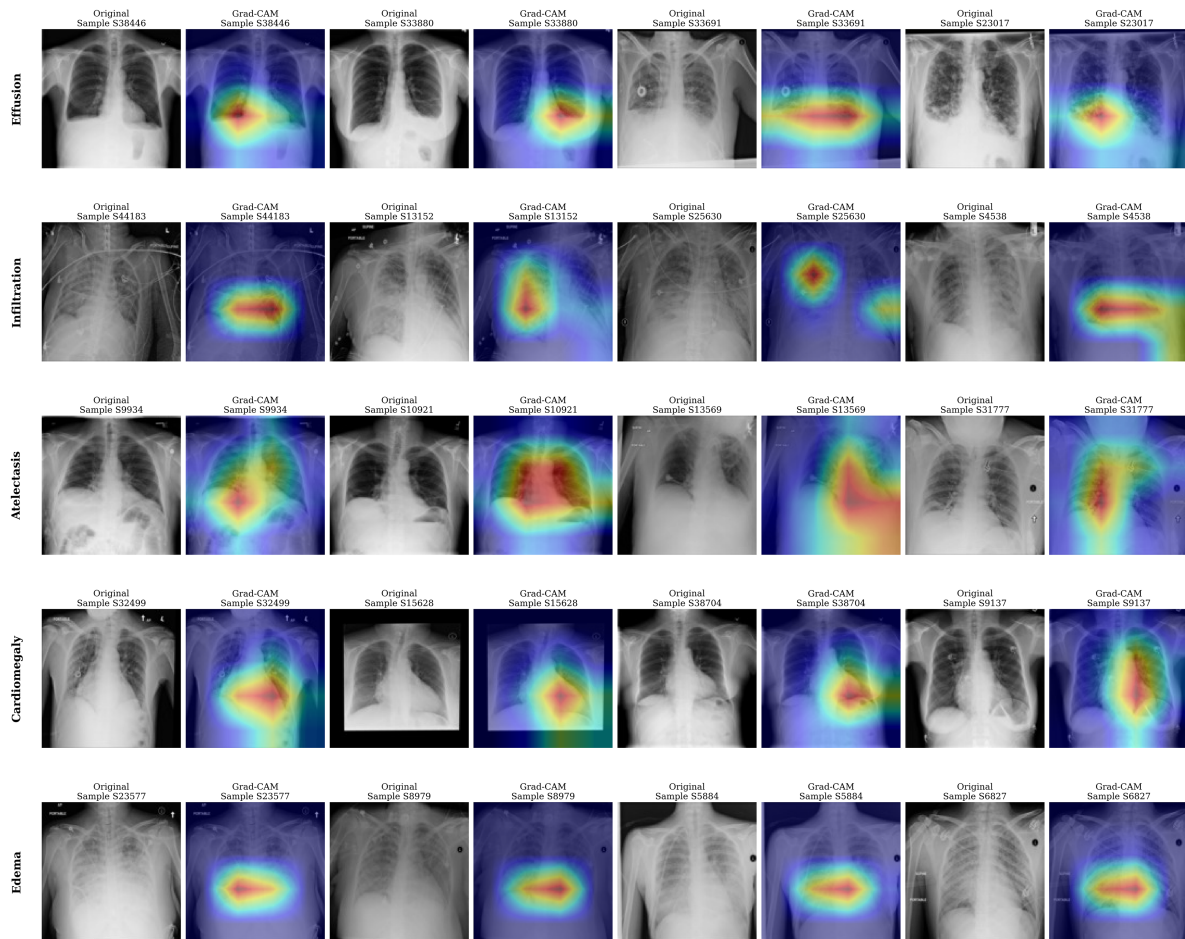


Figure 6: Disease-specific Grad-CAM visualizations obtained from the independently trained NIH model on representative NIH Chest X-rays test samples. The visualized disease categories follow the NIH Chest X-rays label space. For each representative finding, the original chest X-ray and the corresponding Grad-CAM overlay are shown in pairs. The heatmaps highlight disease-related regions contributing to the prediction, including pleural/lower-lung areas for Effusion, diffuse lung-field responses for Infiltration, basal or band-like regions for Atelectasis, cardiac-border regions for Cardiomegaly, and lower-lung opacity patterns for Edema.

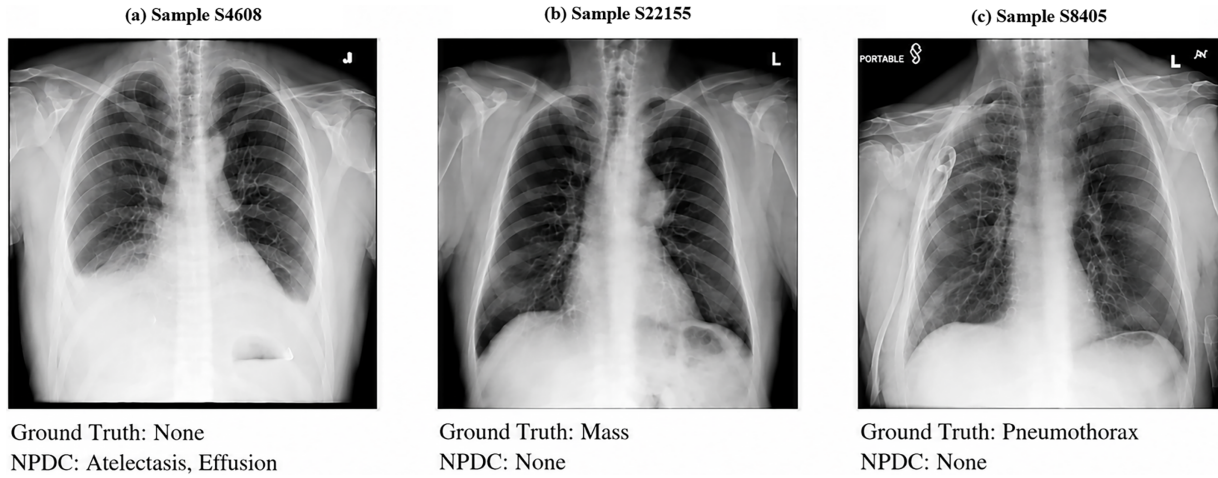


Figure 7: Representative failure cases of proposed method. These cases reflect the difficulty of weakly supervised chest X-ray multi-label prediction under incomplete labels, subtle localized findings, and visually ambiguous radiographic patterns.

3.5 MRCA Component Ablation

To disentangle the effect of different relational components in MRCA, we conduct an ablation study on the CheXpert validation split under the fixed-threshold setting. This setting excludes the influence of decision calibration and allows the performance variations to be attributed primarily to the cross-modal alignment objective. The analysis focuses on the role of the multi-relational target in representation learning, particularly whether incorporating disease-label similarity and tabular phenotype similarity provides additional benefits beyond conventional identity-based image–tabular contrastive supervision.

As shown in Table 5, the ITC-style baseline represents conventional identity-based contrastive alignment, where only the same-patient image–tabular pair is treated as positive and all other samples in the mini-batch are regarded as negatives. Building upon this baseline, MRCA introduces additional clinical relations into the alignment target by incorporating disease-label similarity and tabular phenotype similarity. To isolate their individual effects, we further evaluate two variants, namely “MRCA w/o label similarity” and “MRCA w/o phenotype similarity”, which remove the corresponding relation term from the multi-relational target.

Table 5: Ablation study of MRCA components on the CheXpert validation split. The fixed threshold of 0.5 is used for threshold-dependent metrics. Results are reported as mean $\pm$ standard deviation across repeated runs. The best and second-best mean values for each metric are highlighted in bold and underlined, respectively.

Method	Exact-Match	Macro-F1	Macro-AUROC	Macro-AUPRC	Micro-F1
ITC-style baseline	0.1038 $\pm$ 0.0090	0.1870 $\pm$ 0.0190	0.7695 $\pm$ 0.0079	0.3954 $\pm$ 0.0173	0.4076 $\pm$ 0.0066
MRCA w/o label similarity	0.0868 $\pm$ 0.0052	<u>0.2213</u> $\pm$ 0.0044	0.8081 $\pm$ 0.0175	0.4476 $\pm$ 0.0197	0.4498 $\pm$ 0.0106
MRCA w/o phenotype similarity	0.0788 $\pm$ 0.0046	0.2047 $\pm$ 0.0151	<u>0.8245</u> $\pm$ 0.0063	<u>0.4515</u> $\pm$ 0.0138	0.4657 $\pm$ 0.0385
Full MRCA	<u>0.0938</u> $\pm$ 0.0091	0.2314 $\pm$ 0.0112	0.8335 $\pm$ 0.0027	0.4827 $\pm$ 0.0084	<u>0.4634</u> $\pm$ 0.0420

The full MRCA achieves the highest Macro-F1, Macro-AUROC, and Macro-AUPRC, improving these metrics from 0.1870, 0.7695, and 0.3954 for the ITC-style baseline to 0.2314, 0.8335, and 0.4827, respectively. The relation-specific variants show different metric profiles: removing phenotype similarity yields the highest

Micro-F1, whereas the ITC-style baseline retains the highest Exact Match. These results indicate that disease-label similarity and tabular phenotype similarity contribute differently to the learned relation structure, while their joint use provides the strongest disease-wise discrimination and Macro-F1. Disease-label similarity supplies explicit supervision from shared annotations, whereas tabular phenotype similarity captures the relational geometry induced by the auxiliary tabular encoder. Together, they preserve same-patient pairing while reducing inappropriate repulsion among clinically related non-paired samples.

To further characterize the source of this complementarity, we examine whether the encoder-induced relation differs from explicit disease-label similarity and whether its contribution depends on learned auxiliary representations. Supplementary Table S3 quantifies the agreement between explicit disease-label similarity and encoder-induced tabular phenotype similarity, while Supplementary Table S4 reports controlled experiments using shuffled embeddings and a randomly initialized frozen auxiliary encoder.

3.6 Decision Calibration Ablation

To evaluate the effectiveness of the proposed decision calibration strategy, an ablation study is conducted on the CheXpert validation split using the checkpoint obtained from the proposed MRCA-based pre-training. This experiment is designed to examine whether the improvement in multi-label prediction comes from sample-adaptive threshold estimation, neighborhood-prior refinement, or their combination. Specifically, four decision strategies are compared: fixed thresholding, neighborhood-prior calibration only, sample-adaptive thresholding only, and the full Neighborhood-Prior Decision Calibration (NPDC).

As shown in Table 6, fixed thresholding with a threshold of 0.5 provides the standard decision baseline. However, it yields relatively low Macro-F1, indicating that a uniform threshold is insufficient for imbalanced medical multi-label prediction. This result supports the need for label- and sample-specific decision calibration.

Table 6: Ablation study of decision calibration strategies on the proposed MRCA checkpoint using the CheXpert validation split. Results are reported as mean $\pm$ standard deviation across repeated runs. The best and second-best mean values for each threshold-dependent metric are highlighted in bold and underlined, respectively.

Decision Method	Exact-Match	Macro-AUROC	Macro-AUPRC	Macro-F1	Micro-F1
Fixed 0.5	0.0938 $\pm$ 0.0091	0.8335 $\pm$ 0.0027	0.4827 $\pm$ 0.0084	0.2314 $\pm$ 0.0112	0.4634 $\pm$ 0.0420
Neighborhood-prior only	0.0754 $\pm$ 0.0098	0.8335 $\pm$ 0.0027	0.4827 $\pm$ 0.0084	0.4176 $\pm$ 0.0106	0.4527 $\pm$ 0.0173
Sample-adaptive only	<u>0.1173</u> $\pm$ 0.0014	0.8335 $\pm$ 0.0027	0.4827 $\pm$ 0.0084	0.3828 $\pm$ 0.0168	<u>0.5393</u> $\pm$ 0.0038
Full NPDC	0.1373 $\pm$ 0.0193	0.8335 $\pm$ 0.0027	0.4827 $\pm$ 0.0084	<u>0.4103</u> $\pm$ 0.0061	0.5517 $\pm$ 0.0044

Neighborhood-prior calibration produces the highest Macro-F1 of 0.4176, showing that local disease priors provide effective label-wise correction under imbalanced disease distributions. Sample-adaptive thresholding instead raises Exact Match and Micro-F1 to 0.1173 and 0.5393, respectively, reflecting its effect on sample-specific decision boundaries.

Combining the two components, full NPDC achieves the highest Exact Match of 0.1373 and Micro-F1 of 0.5517, while maintaining a Macro-F1 of 0.4103. This pattern indicates that neighborhood-prior refinement and sample-adaptive threshold estimation influence complementary aspects of score-to-label conversion, yielding a balanced decision profile across the threshold-dependent metrics.

Since all calibration strategies are applied to the same prediction scores, Macro-AUROC and Macro-AUPRC remain unchanged across different decision methods. These metrics evaluate score-ranking quality,

whereas the compared calibration strategies mainly affect the conversion from continuous probabilities to binary predictions.

3.7 Hyper-Parameter Analysis

All hyper-parameter analyses are conducted on the validation set. The test set is used only for the final evaluation and is not involved in hyper-parameter selection. This protocol ensures that the reported sensitivity analyses reflect model selection behavior without introducing test-set leakage.

3.7.1 Effect of MRCA Auxiliary Relation Strengths

The clinical relation matrix in MRCA is constructed by combining paired identity, disease-label similarity, and phenotype similarity. In this analysis, the identity weight is fixed as $\alpha = 1.0$ to preserve the same-patient image–tabular pairing as the dominant alignment anchor. The auxiliary relation strengths β and γ are varied to examine the effects of disease-label similarity and phenotype similarity on representation learning.

It should be noted that α , β , and γ are not constrained to sum to one in our implementation. Therefore, this experiment evaluates the sensitivity to the absolute strengths of the auxiliary clinical relations rather than their normalized proportions. This setting allows us to examine how strongly disease-label overlap and phenotype similarity should contribute to the relational alignment target while keeping identity-based pairing as a stable reference.

As shown in Table 7, introducing auxiliary clinical relations generally improves performance compared with the setting $\beta = 0.0$, $\gamma = 0.0$, which only relies on identity-based pairing. Different combinations favor different evaluation metrics. For example, $\beta = 0.7$, $\gamma = 0.1$ achieves the best Exact Match and Macro-F1, while $\beta = 0.7$, $\gamma = 0.5$ obtains the highest Micro-F1. The setting $\beta = 0.1$, $\gamma = 0.7$ provides competitive performance among phenotype-emphasized configurations, indicating that phenotype similarity can serve as a useful complementary cue for MRCA. However, overly large auxiliary strengths do not consistently improve all metrics, suggesting that excessive reliance on non-identity relations may introduce noisy positive associations. These results support the design choice that same-patient identity should remain the primary alignment anchor, while disease-label and phenotype relations should be incorporated as auxiliary clinical cues.

Table 7: Sensitivity analysis of auxiliary relation strengths β and γ in MRCA, with the identity weight fixed as $\alpha = 1.0$. All results are reported on the validation set. The best result for each metric is highlighted in bold.

β	γ	Exact-Match	Macro-F1	Macro-AUPRC	Micro-F1
0.0	0.0	0.0868	0.1830	0.4045	0.4365
0.0	0.5	0.0838	0.1972	0.4122	0.4369
0.1	0.1	0.1003	0.1887	0.4113	0.4266
0.1	0.3	0.0704	0.1957	0.3630	0.4207
0.1	0.7	0.1063	0.2011	0.4055	0.4464
0.3	0.7	0.1108	0.1758	0.4192	0.3567
0.5	0.1	0.0808	0.1592	0.4216	0.4067
0.7	0.1	0.1168	0.2034	0.4024	0.4202
0.7	0.5	0.0823	0.1938	0.3945	0.4563
1.0	0.7	0.0988	0.1741	0.4180	0.3957

3.7.2 Effect of Top-K Soft Positive Selection

The top- K soft positive selection controls how many clinically related samples are retained when constructing the relational target distribution in MRCA. This analysis is conducted to examine whether a sparse relational positive set is necessary for effective image–tabular alignment. A very small K may discard clinically informative positive samples, whereas a very large K may introduce weakly related or noisy samples into the soft target distribution. The “No top- K ” setting removes this sparsification step and uses all samples in the mini-batch when constructing the relational target.

As shown in Table 8, the performance varies noticeably with different values of K . When K is too small, such as $K = 1$ or $K = 3$, the model obtains relatively low Macro-F1, suggesting that only using very few relational positives is insufficient to capture broader clinical relationships between patients. Increasing K improves the label-wise decision performance, and the best Macro-F1 and Micro-F1 are achieved at $K = 10$. This indicates that retaining a moderate number of clinically related soft positives provides useful relational supervision for multi-label prediction.

Table 8: Sensitivity analysis of top- K soft positive selection in MRCA. All results are reported on the validation set. The best result for each metric is highlighted in bold.

Top- K	Exact-Match	Macro-F1	Macro-AUPRC	Micro-F1
1	0.0853	0.1674	0.3799	0.4085
3	0.0659	0.1606	0.4024	0.4181
5	0.0793	0.1815	0.4087	0.4201
7	0.0808	0.1806	0.3908	0.4150
10	0.0973	0.1986	0.4054	0.4391
15	0.0808	0.1619	0.3924	0.3890
No top- K	0.1138	0.1595	0.3975	0.3567

However, further increasing K to 15 leads to a clear performance drop, especially in Micro-F1. This suggests that excessive relational positives may introduce noisy or weak associations, which can weaken the quality of cross-modal alignment. The “No top- K ” setting achieves the highest Exact Match but substantially lower Macro-F1 and Micro-F1, indicating that using all mini-batch samples as relational positives does not consistently improve label-wise prediction quality. Overall, these results demonstrate the importance of Top- K sparse positive selection in MRCA, and $K = 10$ provides the best balance between preserving clinically useful relations and suppressing noisy positives.

3.7.3 Effect of NPDC Neighborhood Parameters

The neighborhood-prior calibration module introduces two key hyper-parameters: the neighborhood size K_n and the refinement strength λ_n . This analysis is conducted to evaluate how the quality of local disease prior estimation and the strength of threshold refinement affect the final binary multi-label predictions. The neighborhood size K_n controls how many training samples are retrieved from the memory bank to estimate the local disease prior, while λ_n determines how strongly the local–global prior discrepancy adjusts the sample-adaptive threshold.

As shown in Table 9, using a moderate neighborhood size generally leads to stable performance. When K_n is too small, the estimated local disease prior may be sensitive to individual neighbors. When K_n becomes too large, weakly related samples may be included, which can dilute the local disease evidence. The results

show that $K_n = 10$ and $K_n = 20$ achieve comparable performance, while $K_n = 10$ provides a favorable balance between prediction performance and neighborhood retrieval cost.

Table 9: Sensitivity analysis of NPDC neighborhood size K_n and refinement strength λ_n . All results are reported on the validation set. The best and second-best mean values for each metric are highlighted in bold and underlined, respectively.

K_n	λ_n	Exact-Match	Macro-F1	Macro-Recall	Micro-F1
3	0.05	0.1213	0.3752	0.4790	0.5627
5	0.05	0.1228	0.3762	0.4803	0.5648
10	0.05	0.1243	0.3766	0.4805	0.5645
20	0.05	0.1213	0.3749	0.4793	0.5626
10	0.10	0.1243	0.3758	0.4814	0.5650
10	0.20	0.1272	0.3690	0.4753	0.5620
10	0.30	0.1257	0.3660	0.4807	0.5557
10	0.50	0.1243	0.3621	0.4797	0.5481
20	0.10	<u>0.1243</u>	0.3767	<u>0.4821</u>	0.5650
20	0.30	0.1243	0.3728	0.4878	0.5636

The refinement strength λ_n also has a clear influence on the calibration results. A small value such as $\lambda_n = 0.05$ provides stable improvements, whereas larger values gradually reduce Macro-F1 and Micro-F1. This suggests that overly strong neighborhood-prior adjustment may over-correct the adaptive thresholds and introduce noisy prior effects. Although several settings obtain slightly better values on individual metrics, $K_n = 10$ and $\lambda_n = 0.05$ achieve consistently competitive performance across Exact Match, Macro-F1, Macro-Recall, and Micro-F1. Therefore, this setting is adopted as the final NPDC configuration.

Overall, the results indicate that neighborhood-prior decision calibration is not highly sensitive to small changes in K_n , but requires a carefully controlled refinement strength. A moderate neighborhood size and a small calibration strength allow NPDC to exploit local disease priors while avoiding excessive threshold perturbation.

4 Conclusion

In this paper, we investigated chest X-ray multi-label prediction with structured disease-label-guided representation learning, where report-derived tabular disease-label fields provide auxiliary relational guidance during pre-training. To address the limitations of conventional pre-training and fixed-threshold multi-label inference, we proposed a two-stage framework that combines Multi-Relational Cross-Modal Alignment (MRCA) with Neighborhood-Prior Decision Calibration (NPDC).

In the pre-training stage, MRCA replaces identity-only contrastive supervision with soft alignment targets constructed from patient identity, disease-label similarity, and tabular phenotype similarity. This design preserves reliable same-patient image-tabular correspondence while reducing inappropriate repulsion between clinically related non-paired patients. As a result, the image encoder can learn more disease-aware and clinically structured representations. In the prediction stage, the auxiliary tabular branch is removed, and the pre-trained image encoder is transferred to image-only multi-label disease prediction. To improve score-to-label conversion, NPDC estimates sample-label-specific thresholds and further refines them using local neighborhood disease priors retrieved from an embedding-based training memory bank.

Experiments using CheXpert development data with held-out CheXlocalize testing, together with independent experiments on NIH Chest X-rays, demonstrate the effectiveness of the proposed framework. MRCA consistently improves representation quality compared with representative pre-training baselines, including ITC, TIP, MMCL, and PMRL. The decision ablation further demonstrates that sample-adaptive threshold estimation and neighborhood-prior refinement address complementary aspects of score-to-label conversion, improving complete-label consistency and aggregate binary prediction while preserving strong disease-wise performance. The MRCA component ablation also confirms that disease-label similarity and phenotype similarity provide complementary relational cues for cross-modal alignment.

Overall, the proposed framework provides an effective solution for transferring structured disease-label relations through an auxiliary tabular encoder into an image-only chest X-ray prediction model and for improving multi-label decision quality under imbalanced disease distributions. Further evaluation on larger external cohorts, more fine-grained clinical attributes, and localization-aware supervision remains important for assessing the broader applicability of the framework. Within this line of research, extending the current binary-label formulation to uncertainty-preserving disease-label distributions constitutes the next methodological question. Such an extension concerns probabilistic or soft-label representations, uncertainty-specific supervision, and relation measures that preserve the diagnostic ambiguity of uncertain and unmentioned findings.

Acknowledgement: Not applicable.

Funding Statement: This work is supported by the National Natural Science Foundation of China under Grant 62273155, Science and Technology Projects in Guangzhou (2025B01J3018), Science and Technology Project of Ganzhou (2023LNS27051).

Author Contributions: Yuxin Zhang: Conceptualization, Methodology, Writing original draft, Software, Review & Editing. Bin Li: Conceptualization, Funding acquisition, Investigation, Methodology, Writing—Review & Editing, Project administration, Validation. Riqiang Liao: Conceptualization, Data curation, Investigation, Resources, Validation. Lianfang Tian: Investigation, Resources. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The datasets used in this study are publicly available benchmark chest X-ray datasets. The CheXpert dataset was used for MRCA pre-training, downstream fine-tuning, internal validation, and model selection. It is publicly available from the Stanford ML Group and Stanford AIMI repository: <https://stanfordmlgroup.github.io/competitions/chexpert/> (accessed on 01 June 2026). The CheXlocalize dataset was used as a held-out expert-annotated evaluation dataset. It is publicly available from Stanford AIMI: <https://aimi.stanford.edu/datasets/chexlocalize> (accessed on 01 June 2026). The NIH Chest X-rays dataset, also known as ChestX-ray14, was used for independent training, validation, and testing on an additional dataset. It is publicly available from the NIH Clinical Center download site: <https://nihcc.app.box.com/v/ChestXray-NIHCC> (accessed on 01 June 2026). The processed data splits and derived metadata generated during this study are available from the corresponding author upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Supplementary Materials: The supplementary material is available online at <https://www.techscience.com/doi/10.32604/cmes.2026.087646/s1>. Supplementary Table S1 reports the label distributions of CheXpert and CheXlocalize; Supplementary Table S2 reports the label distribution of NIH Chest X-rays; Supplementary Table S3 reports the relation-level similarity analysis; and Supplementary Table S4 reports the controlled auxiliary-encoder experiments.

References

1. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 2097–106.
2. Qin C, Yao D, Shi Y, Song Z. Computer-aided detection in chest radiography based on artificial intelligence: a survey. *Biomed Eng Online*. 2018;17(1):113. doi:10.1186/s12938-018-0544-y.
3. Jiang JP, Ye HJ, Wang L, Yang Y, Jiang Y, Zhan DC. Tabular insights, visual impacts: transferring expertise from tables to images. In: Proceedings of the 41st International Conference on Machine Learning; 2024 Jul 21–27; Vienna, Austria. p. 21988–2009.
4. Holland R, Leingang O, Bogunović H, Riedl S, Fritsche L, Prevost T, et al. Metadata-enhanced contrastive learning from retinal optical coherence tomography images. *Med Image Anal*. 2024;97:103296. doi:10.1016/j.media.2024.103296.
5. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: Meila M, Zhang T, editors. Proceedings of the 38th International Conference on Machine Learning; 2021 Jul 18–24; Virtual. p. 8748–63.
6. Zhang Y, Jiang H, Miura Y, Manning CD, Langlotz CP. Contrastive learning of medical visual representations from paired images and text. In: Proceedings of the 7th Machine Learning for Healthcare Conference; 2022 Aug 5–6; Durham, NC, USA. p. 2–25.
7. Huang SC, Shen L, Lungren MP, Yeung S. GLoRIA: a multimodal global-local representation learning framework for label-efficient medical image recognition. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 3922–31.
8. Bannur S, Hyland S, Liu Q, Pérez-García F, Ilse M, Castro DC, et al. Learning to exploit temporal structure for biomedical vision-language processing. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 15016–27.
9. Wang Z, Wu Z, Agarwal D, Sun J. Medclip: contrastive learning from unpaired medical images and text. *Proc Conf Empir Methods Nat Lang Process*. 2022;2022:3876–87.
10. Hager P, Menten MJ, Rueckert D. Best of both worlds: multimodal contrastive learning with tabular and imaging data. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Los Alamitos, CA, USA. p. 23924–35.
11. Du S, Zheng S, Wang Y, Bai W, O'Regan DP, Qin C. TIP: tabular-image pre-training for multimodal classification with incomplete data. In: Leonardi A, Ricci E, Roth S, Russakovsky O, Sattler T, Varol G, editors. Computer vision-ECCV 2024. Cham, Switzerland: Springer Nature; 2025. p. 478–96.
12. Liu X, Xia X, Ng SK, Chua TS. Principled multimodal representation learning. *IEEE Trans Pattern Anal Mach Intell*. 2026;48(8):9114–28. doi:10.1109/tpami.2026.3675685.
13. Huynh T, Kornblith S, Walter MR, Maire M, Khademi M. Boosting contrastive self-supervised learning with false negative cancellation. In: Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2022 Jun 3–8; Waikoloa, HI, USA. p. 2785–95.
14. Chen TS, Hung WC, Tseng HY, Chien SY, Yang MH. Incremental false negative detection for contrastive learning. In: Proceedings of the Tenth International Conference on Learning Representations; 2022 Apr 25; Virtual.
15. Zhang P, Wu M. Multi-label supervised contrastive learning. *Proc AAAI Conf Artif Intell*. 2024;38:16786–93. doi:10.1609/aaai.v38i15.29619.
16. Khosla P, Teterwak P, Wang C, Sarna A, Tian Y, Isola P, et al. Supervised contrastive learning. In: Proceedings of the 34th International Conference on Neural Information Processing Systems; 2020 Dec 6–12; Vancouver, BC, Canada. p. 18661–73.
17. Chuang CY, Robinson J, Lin YC, Torralba A, Jegelka S. Debaised contrastive learning. In: Proceedings of the 34th International Conference on Neural Information Processing Systems; 2020 Dec 6–12; Vancouver, BC, Canada. p. 8765–75.

18. Holste G, Zhou Y, Wang S, Jaiswal A, Lin M, Zhuge S, et al. Towards long-tailed, multi-label disease classification from chest X-ray: overview of the CXR-LT challenge. *Med Image Anal.* 2024;97:103224. doi:10.1016/j.media.2024.103224.
19. Ge Z, Mahapatra D, Sedai S, Garnavi R, Chakravorty R. Chest x-rays classification: a multi-label and fine-grained problem. *arXiv:1807.07247*. 2018.
20. Lin YJ, Lin CJ. On the thresholding strategy for infrequent labels in multi-label classification. In: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*; 2023 Oct 21–25; Birmingham, UK. New York, NY, USA: Association for Computing Machinery; 2023. p. 1441–50.
21. Pillai I, Fumera G, Roli F. Threshold optimisation for multi-label classifiers. *Pattern Recognit.* 2013;46(7):2055–65. doi:10.1016/j.patcog.2013.01.012.
22. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: *Proceedings of the 34th International Conference on Machine Learning*; 2017 Aug 6–11; Sydney, NSW, Australia. p. 1321–30.
23. Müller R, Kornblith S, Hinton GE. When does label smoothing help? In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*; 2019 Dec 8–14; Vancouver, BC, Canada. p. 4694–703.
24. Menon AK, Jayasumana S, Rawat AS, Jain H, Veit A, Kumar S. Long-tail learning via logit adjustment. In: *Proceedings of the 2021 International Conference on Learning Representations*; 2021 May 4; Vienna, Austria.
25. Wu T, Huang Q, Liu Z, Wang Y, Lin D. Distribution-balanced loss for multi-label classification in long-tailed datasets. In: Vedaldi A, Bischof H, Brox T, Frahm JM, editors. *Computer vision–ECCV 2020*. Cham, Switzerland: Springer International Publishing; 2020. p. 162–78.
26. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, et al. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. *Proc AAAI Conf Artif Intell.* 2019;33(1):590–7.
27. Saporta A, Gui X, Agrawal A, Pareek A, Truong SQH, Nguyen CDT, et al. Benchmarking saliency methods for chest X-ray interpretation. *Nat Mach Intell.* 2022;4(10):867–78. doi:10.1038/s42256-022-00536-x.