CLIP-XRad: Learning Multimodal Representations of Chest X-rays through Contrastive Pretraining with Medical Concept Alignment

Tudor-Octavian Mihăiţă
Department of Computer Science, Babeş-Bolyai University
Cluj-Napoca, Romania
tudor.mihaita@stud.ubbcluj.ro

***Abstract*—Contrastive vision-language pretraining has shown strong performance in learning rich multimodal representations from image-text pairs. However, applying this paradigm to the medical domain presents significant challenges due to the complex visual patterns of medical conditions and the subtle, often ambiguous language used in radiology reports. To address these limitations, we propose CLIP-XRad, a weakly-supervised contrastive vision-language model specialized in chest X-ray interpretation. Our method introduces medical concept alignment into the contrastive learning process via a custom training objective that incorporates both instance-level and concept-level similarity, guiding the model to structure its embedding space around shared clinical semantics. Through extensive experiments, we demonstrate that CLIP-XRad achieves competitive performance in image-text retrieval, while also transferring effectively to downstream tasks such as multi-label classification and report generation. These results demonstrate that concept-aware contrastive pretraining can produce generalizable, semantically meaningful medical representations that provide a data-efficient solution for clinical applications.**

***Index Terms*—Chest X-ray, Contrastive Learning, Vision-Language Pretraining, Cross-Modal Retrieval, Transfer Learning**

I. Introduction

Medical imaging has become a critical component of modern healthcare, enabling clinicians to non-invasively examine internal anatomical structures for the purpose of identifying, monitoring and treating a wide range of pathological conditions. Among the various specialized imaging modalities, chest X-rays (CXR) stand out as the most widely used for diagnosing thoracic diseases, due to their accessibility and diagnostic relevance. Each CXR study is typically accompanied by a radiology report, consisting in a structured free-text description written by medical experts that summarizes key findings and general impressions about the patient's condition. Given the rapid growth in both the usage and significance of medical imaging data, recent research has increasingly focused on developing deep learning methods to support the clinical decision-making process. This scenario can be formally framed as a multimodal learning task, where the objective is to align visual features from medical images with the semantic information encoded in clinical text.

Recent advances in Vision-Language Pretraining (VLP) have shown great promise in learning joint representations from large-scale image-text datasets without any task-specific supervision. Contrastive Language–Image Pretraining (CLIP) [1], in particular, aligns paired samples using a contrastive objective based on cosine similarity and shows strong generalization across a wide range of downstream tasks. Bootstrapping Language-Image Pretraining (BLIP) [2] builds on this paradigm by introducing a unified framework for both vision-language understanding and image-conditioned text generation within a single framework, further demonstrating the adaptability of VLP models across diverse multimodal applications.

Despite their success in general-domain tasks, these models face limitations in medical contexts due to the scarcity of high-quality paired data and the complexity of clinical language and imaging. Standard contrastive objectives assume that only exact pairs are semantically aligned, neglecting shared findings across patient studies and potentially misguiding training.

In this paper, we introduce CLIP-XRad, a weakly-supervised, contrastive Vision-Language Model (VLM) specialized for medical-image text alignment, addressing key limitations of general-purpose approaches for clinical settings. Our method advances contrastive pretraining through a custom loss function, Multi-view Concept Similarity Loss (MCSL), which augments the alignment process by balancing pairwise supervision with concept-level similarity. This enables the model to capture richer clinical semantics beyond individual sample matching. We further investigate how this objective improves multimodal representation learning and supports effective transfer to related medical tasks, including image classification and radiology report generation.

The purpose of this work is centered around answering the following research questions:

RQ1. To what extent does multi-view supervision combined with semantic concept alignment enhance the performance of contrastive vision-language models in learning rich, domain-adapted representations of medical data?

RQ2. Can the learned semantic space during contrastive pretraining be effectively adapted to several tasks in the medical domain, including image classification and report generation?

The paper is organized as follows. Section II reviews related work on Vision-Language Pretraining and its prior applications to the medical domain. Section III presents our proposed methodology in detail, including model architecture, contrastive loss formulation, and strategies for leveraging the pretrained model on downstream tasks. Section IV describes the experimental results and comparisons with existing approaches. Finally, Section V summarizes our contributions and outlines future research directions.

II. Related Work

Vision-text representation learning has shown remarkable success in general-domain settings, where contrastive pretraining enables models to align visual features with corresponding textual descriptions using large-scale paired datasets [1], [3]. However, applying this paradigm to the medical domain is non-trivial and presents several challenges, including limited labeled data, complex domain-specific semantics, and the need for expert-curated annotations. These limitations reduce the model's ability to effectively align imaging studies with their corresponding reports, as reflected through lower retrieval performance. Unlike general-domain tasks, where text often directly describes visible, detectable objects, medical imaging involves subtle, fine-grained visual features that require a deeper semantic understanding and more complex supervision strategies to capture accurately.

Recent research has proposed domain-specific objectives and supervision mechanisms aimed at capturing clinically meaningful alignments, and nuanced semantic relations across studies. In parallel, building on the success of general-purpose multi-task contrastive models [2], increasing attention has been directed toward repurposing these pretrained representations to other related clinical tasks through transfer learning, demonstrating their broader utility within the medical domain.

A. Concept-aware contrastive alignment

Introduced by Wang et al., MedCLIP [4] addresses limitations of contrastive pretraining in clinical scenarios by shifting from instance-level matching to concept-level alignment. Rather than depending on strict image-text pairings in order to define positive samples, it introduces a label-guided semantic matching loss that uses multi-hot encoding vectors representing medical conditions extracted from images and text reports.

These concept vectors define soft similarity targets based on cosine similarity, enabling the model to align images and texts that share clinical semantics, even in the absence of explicit pairing. This allows MedCLIP to be trained on unpaired datasets and guides the learned embedding space to effectively differentiate medical condition semantics.

The resulting representations perform well on classification tasks and class-based image-text retrieval. However, by removing instance-level supervision, the model captures high-level diagnostic similarity at the cost of finer-grained textual understanding present in radiology reports, limiting its effectiveness in downstream tasks that involve more complex clinical language, including report generation. Moreover, the model's performance is sensitive to errors in concept extraction, such as misinterpreted negations or incomplete entity recognition.

B. Multi-view supervision for chest X-ray interpretation

CXR-CLIP [5] is a Vision-Language Pretraining model specifically designed for chest X-ray interpretation. It improves the discriminative capacity of both image and text encoders by leveraging a combination of image-text and image-label datasets. To address the limited availability of high-quality radiology reports, it generates synthetic reports from diagnostic labels using radiology-specific templates.

Another core contribution of CXR-CLIP is the Multi-View Supervision (MVS) strategy. By training on augmented or alternative versions of each modality, the model improves representation robustness while effectively increasing the diversity of training samples. This approach is further supported by two auxiliary contrastive losses, that enforce intra-modal consistency across image and text variations.

CXR-CLIP demonstrates strong performance on image-text retrieval and shows applicability to image classification. However, its ability to capture broader semantics in the learned multimodal embeddings is limited, due to the lack of domain-specific guidance involved in the alignment process.

C. Downstream task adaptation

While contrastive VLP models are primarily designed for image-text retrieval, recent work has explored how their learned representation spaces can be leveraged for a range of tasks that implicitly rely on cross-modal alignment specialized to medical data.

In classification, methods such as CheXzero [6] utilize CLIP-based encoders for zero-shot prediction of CheXpert labels by prompting the model with condition-specific textual queries, enabling inference without explicit task-specific fine-tuning, and achieving impressive performance.

In the context of radiology report generation, models such as RepsNet [7] integrate contrastive supervision within a conditional encoder-decoder framework, by reformulating the task as Visual Question Answering (VQA). These adaptations highlight the versatility of contrastive pretraining in supporting clinical tasks beyond retrieval.

III. Methodology

Following the ideas introduced in prior Vision-Language Pretraining approaches in the medical domain, we propose CLIP-XRad, a model that enhances multimodal alignment by integrating two complementary forms of supervision that capture underlying semantics of medical concepts based on fixed label sets, while preserving and improving instance-level alignment robustness. This section outlines the core methodology of our approach. We first describe the model architecture and the interaction between its components, followed by the contrastive loss formulation used to align chest X-rays and corresponding reports. Finally, we explore the transferability of the learned shared embedding space for downstream multimodal medical tasks.

A. Architecture overview

As established in prior literature, and in accordance with the multimodal nature of the task, CLIP-XRad is formulated as a Vision-Language Model, comprising of a dual-encoder module, built upon the Transformer architecture [8], as shown in Figure 1.

For the image encoder, we use a Swin Transformer Tiny backbone [9], with a patch size of 4x4 and a shifting window size of 7x7. The input chest X-ray images are resized to 224x224, augmented using medical-specific transformations, and passed through the transformer blocks to produce hierarchical visual representations. The [CLS] token embedding from the final stage is extracted as the global image feature descriptor for vision-language alignment and downstream tasks.

As for the text encoder, we employ ClinicalBERT, representing a BERT [10] pretrained on medical vocabulary, to encode the radiology reports. The resulting textual embeddings are used during pretraining for contrastive alignment with visual embeddings, and later support zero-shot classification by enabling similarity-based matching with encoded label representations.

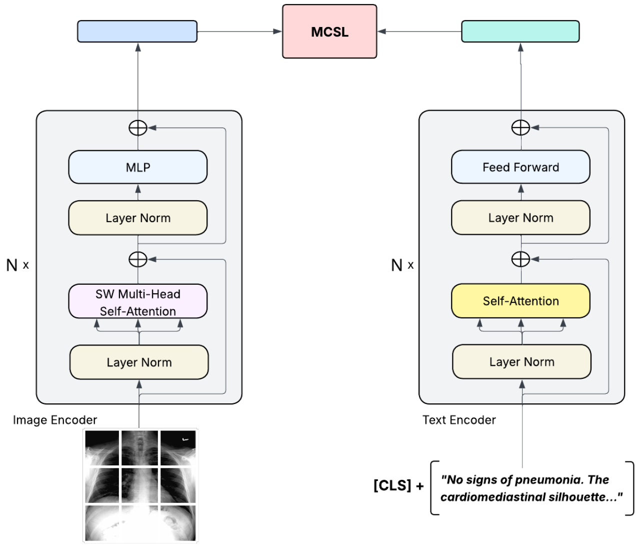

Fig. 1: CLIP-XRad dual-encoder architecture using Swin Transformer for images and ClinicalBERT for text, aligned via the Multi-view Concept Similarity Loss (MCSL).

B. Multi-view Concept Similarity Loss

The goal of Vision-Language Pretraining is to learn a shared embedding space where semantically aligned image-text pairs are pulled closer, while less relevant pairs are pushed apart. This is achieved by optimizing a contrastive loss function over cross-modal representations. To effectively define this objective in the clinical setting, it is essential to establish what constitutes a positive or negative pair within the specific data domain, and to determine an appropriate similarity measure to guide the alignment process.

To address the limitations of strict pairwise matching, especially in multi-label or unpaired datasets, we introduce structured semantic supervision using concept label vectors derived from CheXpert findings [11]. Each sample is associated with a multi-label binary vector defined as $l_i \in \{0,1\}^C$, encoding the presence of $C = 14$ radiological findings.

Instead of relying on a direct dot product between the label vectors to estimate sample similarity, we adopt the Jaccard Index, which quantifies the proportion of shared clinical findings relative to their union. This measure provides a more semantically grounded notion of similarity in the multi-label setting of medical diagnosis and offers greater robustness to sparse positive labels and severe class imbalance, both common characteristics of clinical datasets. The Jaccard similarity function between two samples i and j is given by Formula 1.

$$\text{sim}_{Jaccard}(i,j) = \frac{|l_i \cap l_j|}{|l_i \cup l_j|} = \frac{l_i \cdot l_j}{\|l_i\|_1 + \|l_j\|_1 - l_i \cdot l_j} \tag{1}$$

This results in a similarity matrix $J \in \mathbb{R}^{N\times N}$ over a batch of samples of size N, which is row-wise normalized with a temperature-scaled softmax, excluding self-matches, to produce the soft target matrix $\hat{\mathbf{J}}$ as shown in Formula 2. Following the formulation of the InfoNCE loss in CLIP [1], i denotes the anchor sample, j a candidate from the other modality, and k indexes all candidates used in normalization.

$$\hat{\mathbf{J}}_{i,j} = \frac{\exp(\text{sim}_{Jaccard}(i,j)/\tau)}{\sum_{k=1}^{N} \exp(\text{sim}_{Jaccard}(i,j)/\tau)} \tag{2}$$

To incorporate both semantic and instance-level supervision, we construct the final target distribution as illustrated in Figure 2 by blending the semantic similarity matrix with a one-hot pairing identity matrix $\mathbf{I}$, which enforces alignment between ground-truth image-text pairs. The scalar λ controls the relative importance of semantic supervision, determining the extent to which concept-level alignment guides the representation learning process. To ensure the resulting target distribution remains numerically stable and that both supervision components are proportionally integrated, we normalize the weighted sum by $(1+\lambda)$, helping maintain stable gradient flow during training. The resulting semantic target matrix $\hat{\mathbf{S}}$ is computed as shown in Formula 3.

$$\hat{\mathbf{S}}_{i,j} = \frac{\mathbf{I}_{i,j} + \lambda \cdot \hat{\mathbf{J}}_{i,j}}{1+\lambda} \tag{3}$$

Simultaneously, the logits matrix is computed as the normalized dot-product similarity scores between image and text embeddings e_i^{img} and e_j^{txt}, according to Formula 4.

$$\hat{\mathbf{L}}i,j = \frac{\exp(e_i^{\text{img}} \cdot (e_j^{\text{txt}})^\top/\tau)}{\sum_{k=1}^{N} \exp(e_i^{\text{img}} \cdot (e_k^{\text{txt}})^\top/\tau)} \tag{4}$$

A soft cross-entropy loss is applied between predicted logits and semantic targets in both directions, from image-to-text and from text-to-image, with the latter involving transposed indexing, following the definition in Formula 5.

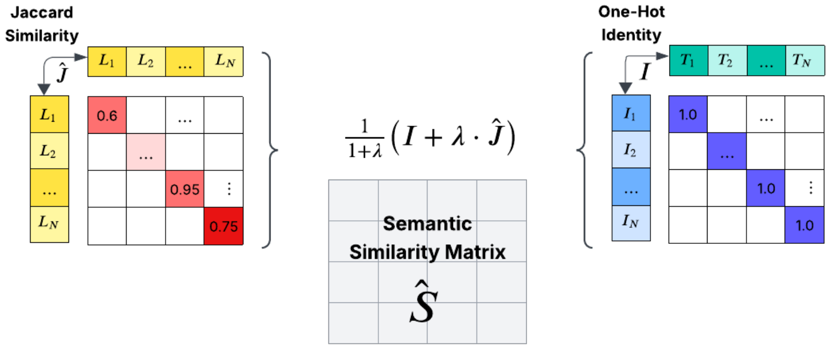

Fig. 2: Construction of the Semantic Similarity Matrix $\hat{\mathbf{S}}$. The final targets are obtained by blending the Jaccard matrix $\hat{\mathbf{J}}$ with the identity matrix $\mathbf{I}$, representing exact image-text pairings. The contribution of medical concept alignment is controlled by the scalar λ, and the resulting sum is further normalized by $(1+\lambda)$.

$$\mathcal{L}^{i\to t}_{semantic} = -\sum_{i=1}^{N}\sum_{j=1}^{N} \hat{\mathbf{S}}_{i,j} \cdot \log \hat{\mathbf{L}}_{i,k},$$
$$\mathcal{L}^{t\to i}_{semantic} = -\sum_{j=1}^{N}\sum_{i=1}^{N} \hat{\mathbf{S}}_{j,i} \cdot \log \hat{\mathbf{L}}_{k,j} \tag{5}$$

The final semantic loss (Formula 6) is obtained by averaging the bidirectional losses, ensuring balanced alignment from both modalities.

$$\mathcal{L}_{semantic} = \frac{1}{2}(\mathcal{L}^{i\to t}_{semantic} + \mathcal{L}^{t\to i}_{semantic}) \tag{6}$$

To improve generalization, we apply this semantic loss under a multi-view supervision (MVS) setup inspired by CXR-CLIP [5], where augmented views I', T' are created. Illustrated in Figure 3, we compute the semantic loss across all combinations of original and augmented views to obtain the final Multi-view Concept Similarity Loss (MCSL) from Formula 7.

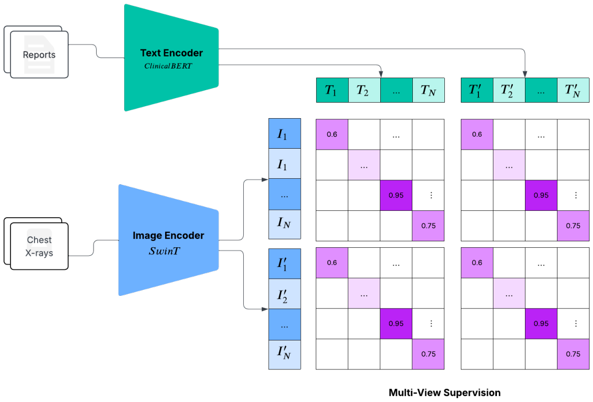

Fig. 3: Multi-View Supervision Strategy used in CLIP-XRad during contrastive pretraining. Each image and text sample is augmented to create four cross-modal pairings.

$$\mathcal{L}_{\text{MCSL}} = \frac{1}{4}\sum_{\substack{i\in\{I,I'\}\\ t\in\{T,T'\}}} \mathcal{L}_{\text{semantic}}(i,t) \tag{7}$$

This objective facilitates alignment not just of paired samples, but of any semantically similar image-text combinations, thereby enhancing the model's ability to capture clinically meaningful patterns within the embedding space. By grounding the alignment in shared medical concepts, the model develops a more fine-grained understanding of pathological findings. Furthermore, since CheXpert labels can be automatically extracted [11], and even refined using neural-enhanced labeling methods such as CheXbert [12], this training setup can extend to unpaired or partially labeled data by potentially generating synthetic, templated reports from concept labels, enabling more robust and flexible pretraining.

C. Zero-shot and supervised image classification

In order to effectively assess the transferability of CLIP-XRad's concept-aware learned representations, we apply the model to multi-label chest X-ray classification, targeting the 14 CheXpert conditions. Following prior conventions [11], uncertain labels (-1) are masked as 0 to focus evaluation on confidently annotated findings, consistent with our pretraining objective that emphasizes clear semantic alignment.

We evaluate two complementary strategies for downstream classification. In zero-shot setting, where the pretrained encoders are leveraged with no additional fine-tuning, classification is formulated as 14 independent binary decisions. For each concept c_i, we define a pair of text prompts describing its presence ($e^{\text{pos}}_{c_i}$) or absence ($e^{\text{neg}}_{c_i}$), and compute cosine similarity between the image embedding $e_{\text{img}} \in \mathbb{R}^d$ and each of the two prompt embeddings. These scores are scaled by a learned temperature parameter and normalized via softmax to result in a probability of presence, mathematically defined as in Formula 8.

$$P(c_i = 1 \mid e_{img}) = \frac{\exp(s_i^{pos}/\tau)}{\exp(s_i^{pos}/\tau) + \exp(s_i^{neg}/\tau)} \tag{8}$$

In supervised setting, we attach a lightweight Multi-Layer Perceptron (MLP) to the frozen image encoder and fine-tune it on labeled data. The MLP, defined as a function $f : \mathbb{R}^d \rightarrow \mathbb{R}^C$, where $C = 14$ represents the number of possible label classes, maps the embedding e_{img} to class-wise probabilities, via a sigmoid-activated output layer as shown in Formula 9.

$$\hat{c} = \sigma(f(e_{img})) \in [0, 1]^C \tag{9}$$

To convert predicted probabilities into binary decisions, we apply a label-wise thresholding strategy. Rather than using a fixed cut-off (e.g., 0.5), we determine the optimal threshold for each condition by maximizing Youden's J Index on the validation dataset split, as defined in Formula 10. This approach provides a principled way to balance sensitivity and specificity, improving overall classification performance across diverse clinical findings. It is particularly beneficial in the presence of label imbalance and ensures that the decision boundaries align with the semantic patterns captured in the learned embedding space.

$$J = \frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}} - 1 \tag{10}$$

Together, these strategies highlight both the generalization capability of the learned image representations and the model's discriminative effectiveness, crucial in real-world clinical environments.

D. Radiology report generation from visual embeddings

Another key application enabled by the learned multimodal representations is conditional radiology report generation. This task can be inherently framed as an encoder-decoder problem, in which a text decoder generates free-text diagnostic reports by conditioning on the visual features extracted from the pretrained image encoder.

To facilitate this, we propose the adaptation of a ClinicalBERT-based decoder with two architectural modifications inspired by BLIP's unified framework [2]. First, we extend each Transformer block with a Cross-Attention layer inserted after the Self-Attention mechanism, allowing the decoder to attend to projected image embeddings. Second, a language modeling head is attached on top of the frozen encoder to support autoregressive text generation, starting from a special beginning-of-sequence token, `<report>`.

In addition to visual conditioning, we further propose incorporating concept-aware prompts derived from a lightweight classification module built on the same contrastive pretrained backbone. These prompts provide explicit semantic guidance toward clinically relevant entities during generation.

The overall process is summarized in Algorithm 1, where an image is encoded, concept prompts are generated from predicted labels, and a conditioned decoder synthesizes the final report. This formulation supports the efficiency of the contrastive pretraining strategy adopted by CLIP-XRad, which enables downstream cross-modal generation with minimal architectural extensions and targeted fine-tuning.

IV. Experimental evaluation

A. Dataset

We conduct our experiments on the MIMIC-CXR-JPG dataset [13], [14], a large-scale benchmark comprising 377,110 chest X-ray studies paired with corresponding radiology reports. For computational efficiency, we rely on the JPG-compressed version of the original DICOM images, and integrate the CXR-PRO [15] variant of the reports, which removes references to prior studies, ensuring consistent alignment with the current imaging findings. Structured diagnostic labels are provided via the CheXpert labeler, covering 14 common thoracic conditions.

To align with our multimodal, paired training strategy, we filter the dataset to retain 253,993 samples that include both CheXpert annotations and CXR-PRO reports. Following the official dataset split, we preserve the test split (2,900 samples), and we allocate 15% of the training data for validation. This is done using a multi-label stratified shuffle split to preserve the clinical label distribution and ensure robust supervision despite the inherent class imbalance of medical data.

B. Experimental setup

We train our dual-encoder model to produce image and text embeddings that are projected in a shared latent space of dimension 768. The semantic similarity supervision is weighted with a scalar value of $\lambda = 0.7$, favoring concept-level alignment during training. The temperature parameter for scaling the cross-modal embeddings is initialized at 0.07, allowing the model to prioritize hard negatives early in training while learning to adapt this value dynamically. Optimization is performed using the AdamW optimizer [16] with a cosine annealing learning rate schedule, starting at 3×10^{-5}, chosen to balance convergence speed and stability. A batch size of 64 is used to ensure a sufficient number of diverse samples at each training step for robust similarity alignment.

For visual inputs, all X-ray images are resized to 224x224 and augmented during training, including random resized cropping, color jittering, and CLAHE-based contrast enhancement [17], which highlights local clinical structures critical for diagnosis. On the textual side, sentence-swapping augmentation is employed to improve generalization and robustness to variations in report structure. Validation and test sets are processed with minimal transformations to retain clinical realism.

All experiments are conducted on a single NVIDIA Testa A100 GPU 40GB VRAM, using mixed-precision training to maximize computational efficiency and minimize memory consumption.

C. Results and discussion

The performance in multimodal alignment capabilities of CLIP-XRad is effectively evaluated through image-text retrieval, complemented by a qualitative analysis of the learned embedding space using a t-SNE projection. In addition, we assess the model's representation transfer by applying it to multi-label classification as a downstream medical task.

```
Input: X-ray image $X$, Frozen image encoder $f_{\text{img}}$, Image projection $f_{\text{proj}}$, Classification module $f_{\text{cls}}$, Prompt
       constructor $\mathcal{P}$, Text decoder $f_{\text{dec}}$, Tokenizer $\mathcal{T}$
Output: Generated radiology report $\hat{R}$
$v \leftarrow f_{\text{img}}(X)$
$v \leftarrow f_{\text{proj}}(v)$ ;                          /* Project image embeddings to decoder hidden size */
$v \in \mathbb{R}^{B \times D} \rightarrow \tilde{v} \in \mathbb{R}^{B \times 1 \times D}$ ;                                /* Add sequence dimension */
$\hat{y} \leftarrow f_{\text{cls}}(v)$
$p \leftarrow \mathcal{P}(\hat{y})$
$t_{\text{input}} \leftarrow \mathcal{T}(p$ + "<report>") ;                        /* '+' represents concatenation */
for $i = 1$ to $B$ do
    for $j = 1$ to $L_v$ do
        $a_{\text{enc}}[i][j] \leftarrow 1$ ;                   /* Set attention mask to 1 for all positions */
    end
end
$\hat{R} \leftarrow f_{\text{dec}}(t_{\text{input}},\ v,\ a_{\text{enc}})$
return $\hat{R}$
```

Algorithm 1: Conditional report generation adaptation strategy using visual features and concept prompts from the CLIP-XRad pretrained backbone.

1) Image-text retrieval: We report retrieval metrics under two evaluation settings: exact report matching (Recall@K), where only the ground-truth report for a given chest X-ray is considered correct, and class-based retrieval (Precision@K), where a sample is counted as relevant if it shares at least one CheXpert label with the query. While Recall@K captures fine-grained alignment, Precision@K better reflects our multi-label training setup and is more relevant to real-world clinical scenarios. However, this setting makes direct comparison with prior work more challenging. Notably, MedCLIP [4] is evaluated on a simplified single-label subset (CheXpert5x200), whereas our setup uses the full set of 14 CheXpert classes and the official MIMIC-CXR test split.

Model	**Recall@K**			**Precision@K**		
	@1	@5	@10	@1	@5	@10
CXR-CLIP [5]	21.6	48.9	60.2	-	-	-
MedCLIP [4]	1.1	1.4	5.5	45.0	48.0	50.0
CLIP-XRad (ours)	**3.8**	**16.2**	**25.9**	**48.3**	**47.4**	**46.5**

TABLE I: Retrieval performance comparison with related work in both exact image-text retrieval using Recall@K and class-based retrieval using Precision@K.

As shown in Table I, CLIP-XRad achieves a Recall@1 of **3.8%**, outperforming MedCLIP (1.1%), by 2.7 percentage points and exceeding it by over 20% for Recall@10. However, it underperforms compared to CXR-CLIP, which achieves 21.6% for Recall@1. In class-based retrieval evaluation, our model reaches a Precision@1 of **48.3%**, closely matching MedCLIP's 45.0%. Despite differences in evaluation setups, these results suggest that CLIP-XRad provides a strong balance between instance-level alignment and semantic-level similarity.

Overall, the results support the effectiveness of our contrastive objective, which encourages deeper clinical understanding while preserving the ability to retrieve accurate matches.

2) Comparison with baseline models: The performance of our specialized contrastive pretraining approach for medical data is further evidenced by a comparison with foundational Vision-Language Models in Table II under a similar exact pair retrieval setup, where CLIP-XRad consistently outperforms general-purpose models, CLIP [1] and BLIP [2], across all Recall@K metrics. These findings highlight that incorporating domain-specific supervision is crucial for learning clinically meaningful image-text alignments.

To visually support this, a t-SNE projection of the image embeddings is presented in Figure 4. The visualization reveals that CLIP-XRad clusters samples with overlapping pathologies, reflecting its capacity to capture clinical semantics. Still, the variability of medical data and frequent co-occurrence of multiple findings contribute to less clearly defined cluster boundaries.

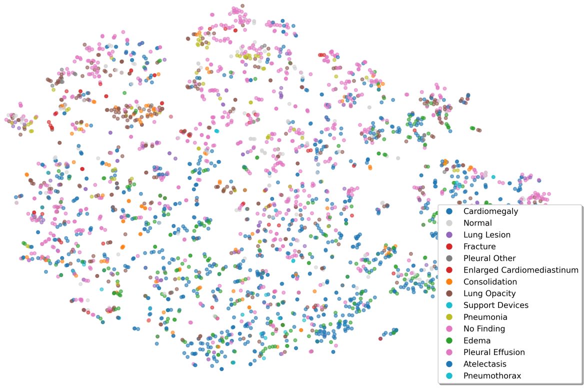

Fig. 4: t-SNE visualization plot of image embeddings from CLIP-XRad, colored by CheXpert pathology labels.

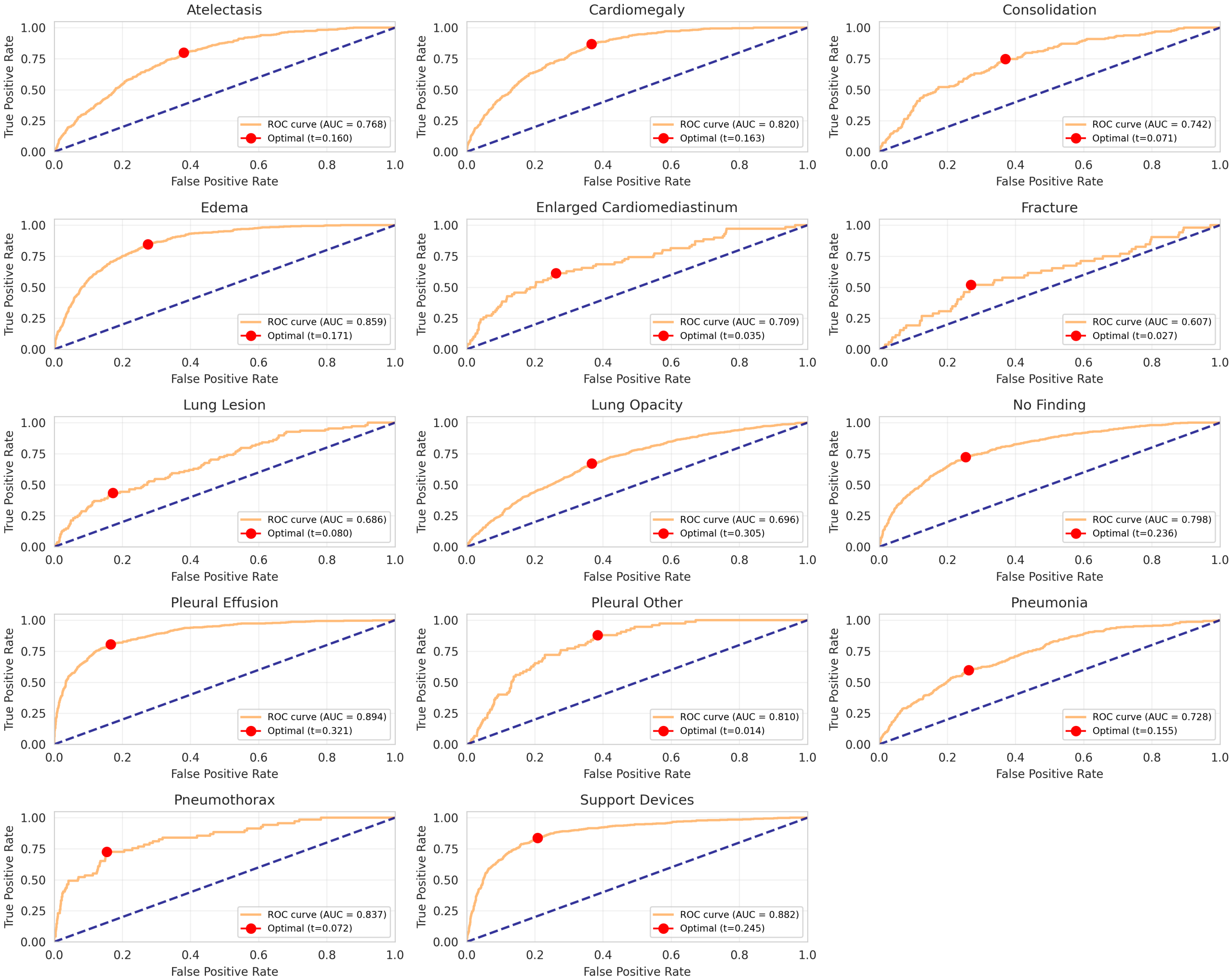

Fig. 5: ROC curves for CheXpert pathology classification with CLIP-XRad. Each subplot presents the AUC score and the optimal decision threshold (highlighted as a red dot) for one of the 14 CheXpert labels.

Model (MIMIC-CXR)	**Recall@1**	**Recall@5**	**Recall@10**
CLIP [1]	0.1	0.3	1.2
BLIP [2]	0.8	4.1	6.8
CLIP-XRad (ours)	**3.8**	**16.2**	**25.9**

TABLE II: Comparison of image-text retrieval performance with baseline contrastive models. All models are fine-tuned on MIMIC-CXR [13].

3) Medical image classification: We evaluate CLIP-XRad on multi-label image classification for the 14 CheXpert pathologies under two settings: zero-shot inference, using similarity-based scoring between image embeddings and textual prompts for each medical concept, and supervised classification, using a fine-tuned classification head on top of the frozen image encoder. For fair comparison with related work, we report Accuracy (ACC) as the primary metric, complemented by Area Under the ROC Curve (AUC) to capture class-wise discriminative performance. Per-label threshold optimization is applied to mitigate class imbalance.

As shown in Table III, CLIP-XRad achieves strong performance in both configurations. In the zero-shot setting, it obtains an accuracy of **66.7%**, outperforming MedCLIP (50.2%, Δ +16.5%) and CXR-CLIP (62.8%, Δ +3.9%). This trend continues in the supervised setup, where CLIP-XRad reaches **72.7%**, again surpassing MedCLIP (56.5%, Δ +16.2%) and CXR-CLIP (65.7%, Δ +7.0%).

Model	**Zero-Shot ACC**	**Supervised ACC**
MedCLIP [4]	50.2	56.5
CXR-CLIP [5]	62.8	65.7
CLIP-XRad (ours)	**66.7**	**72.7**

TABLE III: Comparison of concept classification accuracy under zero-shot and supervised fine-tuning settings.

These results highlight two key observations: first, the consistent performance gains across both evaluation modes confirm the benefit of our concept-aware alignment approach in producing clinically meaningful representations. Second, the strong results in the fine-tuned setting affirm the utility of the learned embeddings as a foundation for downstream medical tasks. To further support this, Figure 5 presents the Receiver Operator Characteristics (ROC) curves per label, illustrating the variability in performance across findings driven by class imbalance and the inherent difficulty of identifying certain conditions.

V. Conclusions and future work

This paper introduced CLIP-XRad, a new approach for contrastive vision-language pretraining adapted to the complexity of the medical domain. The model achieves a **2.7%** improvement in Recall@1 over MedCLIP [4] for exact image-text retrieval, while underperforming compared to CXR-CLIP [5], which is optimized specifically for ground-truth pair matching. In class-based retrieval, measured by Precision@K, CLIP-XRad performs comparably to MedCLIP, though direct comparison is limited by differences in evaluation setups. For downstream adaptation to multi-label classification, the model improves zero-shot accuracy by **16.5%** and supervised accuracy by **16.2%** relative to MedCLIP.

The conducted experiments presented in this work address the research questions outlined in Section I. As an answer to RQ1, our findings demonstrate that the proposed MCSL contrastive objective enhances multimodal representation learning of chest X-rays and corresponding radiology reports by incorporating concept-level supervision. This ensures that relevant condition-specific information is captured and drives efficient retrieval performance for relevant cases, balanced between exact matching capabilities and semantically related cases.

Furthermore, for answering RQ2, we emphasized that the learned representation space exhibits strong transferability to downstream tasks. With minimal adaptation, CLIP-XRad achieved strong performance in multi-label image classification and lays the foundation for potential future extensions to radiology report generation, leveraging the same pretrained backbone under prompt-based guidance.

Our results highlight multiple opportunities for future work. First, building upon MedCLIP's decoupled training approach [4], hybrid setups that combine paired and unpaired datasets could better balance fine-grained matching with broader concept understanding, exposing the model to a greater diversity of medical samples. Another valuable extension would involve enriching the label supervision scheme with additional contextual indicators, such as positional information and uncertainty or severity levels, which could help the model capture more complex medical semantics. While this expands the label space and increases task complexity, it could help the model capture richer medical semantics and enhance interpretability in downstream applications. Moreover, incorporating structured medical knowledge graphs [18] could further enhance semantic grounding and strengthen the model's ability to interpret medical vocabulary within the text modality.

References

[1] A. Radford, J. W. Kim *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of ICML 2021*, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021, pp. 8748–8763. [Online]. Available: http://proceedings.mlr.press/v139/radford21a.html

[2] J. Li, D. Li *et al.*, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," in *Proceedings of ICML*, vol. 162, 2022, pp. 12 888–12 900. [Online]. Available: https://proceedings.mlr.press/v162/li22n.html

[3] C. Jia, Y. Yang *et al.*, "Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision," in *Proceedings of ISML 2021*, vol. 139. Proceedings of Machine Learning Research, 2021, pp. 4904–4916. [Online]. Available: https://proceedings.mlr.press/v139/jia21b.html

[4] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "MedCLIP: Contrastive Learning from Unpaired Medical Images and Text," in *Proceedings of EMNLP 2022*. Abu Dhabi, United Arab Emirates: ACL, Dec. 2022, pp. 3876–3887. [Online]. Available: https://aclanthology.org/2022.emnlp-main.256/

[5] K. You, J. Gu, and thers, "CXR-CLIP: Toward Large Scale Chest X-ray Language-Image Pre-training," *CoRR*, vol. abs/2310.13292, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2310.13292

[6] E. Tiu, E. Talius *et al.*, "Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning," *Nature Biomedical Engineering*, vol. 6, no. 12, pp. 1399–1406, 2022. [Online]. Available: https://doi.org/10.1038/s41551-022-00936-9

[7] A. K. Tanwani, J. K. Barral, and D. Freedman, "RepsNet: Combining Vision with Language for Automated Medical Reports," in *Proceedings of MICCAI 2022*, ser. Lecture Notes in Computer Science, vol. 13435. Springer, 2022, pp. 714–724. [Online]. Available: https://doi.org/10.1007/978-3-031-16443-9_68

[8] A. Vaswani, N. Shazeer, and othres, "Attention is All you Need," in *Proceedings of NIPS'17*. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010. [Online]. Available: https://api.semanticscholar.org/CorpusID:13756489

[9] Z. Liu, Y. Lin *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," *Proceedings of ICCV 2021*, pp. 9992–10 002, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:232352874

[10] J. Devlin, M.-W. Chang *et al.*, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of NAACL 2019*. ACL, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423/

[11] J. Irvin, P. Rajpurkar *et al.*, "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison," *Proceedings of AAAI*, vol. 33, pp. 590–597, 2019.

[12] A. Smit, S. Jain *et al.*, "Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT," in *Proceedings of EMNLP 2020*. ACL, Nov. 2020, pp. 1500–1519. [Online]. Available: https://aclanthology.org/2020.emnlp-main.117/

[13] A. Johnson, T. Pollard *et al.*, "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific Data*, vol. 6, p. 317, 12 2019.

[14] A. Johnson, M. Lungren *et al.*, "MIMIC-CXR-JPG - chest radiographs with structured labels (version 2.1.0)," https://doi.org/10.13026/jsn5-t979, 2024.

[15] V. Ramesh, N. A. Chi, and P. Rajpurkar, "Improving Radiology Report Generation Systems by Removing Hallucinated References to Non-existent Priors," in *Proceedings of ML4H 2022*, vol. 193, 28 Nov 2022, pp. 456–473. [Online]. Available: https://proceedings.mlr.press/v193/ramesh22a.html

[16] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proceedings of ICLR 2019*. OpenReview.net, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[17] E. D. Pisano, S. Zong *et al.*, "Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms," *Journal of Digital Imaging*, vol. 11, no. 4, pp. 193–200, Nov. 1998.

[18] X. Hou, Z. Liu *et al.*, "MKCL: Medical Knowledge with Contrastive Learning model for radiology report generation," *Journal of Biomedical Informatics*, vol. 146, p. 104496, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046423002174