A Cross-Modal Retrieval Framework for Radiology Reports and Chest X-Ray Images

Vijayalaxmi Mekali
Department of Computer Science and Engineering, K. S. Institute of Technology, Bengaluru, India
vijayalaxmimekali@ksit.edu.in

M. P. Sowbhagya
Department of Computer Science and Engineering, K. S. Institute of Technology, Bengaluru, India
sowbhagya.mp@gmail.com

H. D. Aparna
Department of Computer Science and Business Systems, Dayananda Sagar College of Engineering, Bengaluru, India
aparna.havanur@gmail.com

Deepa K. Mathew
Artificial Intelligence and Machine Learning Department, Dayananda Sagar Academy of Technology & Management (DSATM), Bangalore, India
deepa-aiml@dsatm.edu.in

Vandana Singh
Department of Computer Science & Engineering, Amity School of Engineering and Technology, Ranchi, India
vsingh@rnc.amity.edu

Nitesh N Nikam
Department of Electrical Engineering, CSMSS Chh. Shahu College of Engineering, Chhatrapati Sambhajinagar, India
niteshnikam1408@gmail.com

Ganesh B. Dongre
Department of Electronics and Computer Engineering, CSMSS Chh. Shahu College of Engineering, Chhatrapati Sambhajinagar, India
ganeshbdongre@gmail.com

Yogesh H Bhosale
Department of Computer Science & Engineering, CSMSS Chh. Shahu College of Engineering, Kanchanwadi, Chhatrapati Sambhajinagar, India
yogeshbhosale988@gmail.com (corresponding author)

Received: 24 December 2025 | Revised: 8 January 2026 and 31 January 2026 | Accepted: 2 February 2026

 DOI: https://doi.org/10.48084/etasr.17163

ABSTRACT

Accurate retrieval of relevant radiology reports and images is essential for clinical decision support and large-scale medical data management. This study proposes MedFuse-CLIP, a domain-tuned cross-modal retrieval framework that aligns chest X-ray images with corresponding radiology reports using a dual-encoder OpenCLIP ViT-B/32 backbone. The model introduces two key innovations: adaptive semantic

hard-negative mining, which enhances discriminative learning across visually similar pathologies, and a retrieval-aware margin contrastive loss, which stabilizes alignment within the embedding space. Experiments on the MIMIC-CXR-JPG dataset demonstrate strong performance, achieving Recall@10 scores of 87.5% for image→text retrieval and 83.4% for text→image retrieval under fine-tuned cross-modal alignment, along with an average AUROC of 0.879 for zero-shot disease classification across 14 thoracic pathologies. Dimensionality reduction analysis confirmed that compact 256-dimensional embeddings preserve more than 98% of retrieval accuracy while halving storage requirements. The results indicate that MedFuse-CLIP matches or exceeds existing radiology vision-language models such as BioViL and GLoRIA while operating efficiently on a single consumer GPU.

Keywords-radiology retrieval; cross-modal learning; vision–language model; contrastive learning; CLIP; medical imaging AI

I. INTRODUCTION

The exponential growth of digital medical imaging has resulted in the creation of large-scale Chest X-ray (CXR) datasets paired with structured radiology reports, enabling data-driven clinical analysis and retrieval research [1]. Efficient retrieval of relevant cases based on textual or visual queries has become increasingly important for clinical diagnosis, decision support, and retrospective studies. Traditional retrieval systems often rely on metadata or keyword searches, which fail to capture the rich semantic relationships between radiological findings and textual descriptions. The advent of vision language models, such as CLIP, has opened new possibilities for cross-modal retrieval, enabling unified embedding spaces where images and texts are semantically aligned. Recent advances in vision-language representation learning have shown remarkable potential in connecting image and text modalities. Models such as CLIP and ALIGN learn joint embeddings via contrastive learning on large image-text corpora [2]. In the medical domain, frameworks such as BioViL, GLoRIA, and CXR-CLIP have adapted this paradigm to radiology, showing promising results in image-report matching and zero-shot disease classification. Despite these advances, two limitations persist: (i) existing models are trained on mixed or limited datasets that lack comprehensive coverage of clinical diversity, and (ii) their negative sampling strategies often treat unrelated reports as negatives without considering semantic proximity or patient context, leading to suboptimal alignment in diagnostically similar cases. Although current vision language models can map radiology images and reports into a shared space, they still struggle to discriminate subtle visual and linguistic differences among similar diseases and often misalign semantically overlapping but clinically distinct findings. Recent studies have demonstrated the effectiveness of deep learning models for automated chest X-ray analysis, particularly for pneumonia and other thoracic disease detection, highlighting the growing role of data-driven approaches in radiological decision support [3].

The main objective of this study was to design an efficient cross-modal retrieval framework that accurately aligns chest X-ray images and their corresponding radiology reports using a contrastive learning approach. The proposed model adapts the OpenCLIP ViT-B/32 backbone for medical domain alignment through structured disease prompts. It further employs an adaptive hard-negative mining strategy to identify semantically similar yet clinically distinct samples, and a retrieval-aware margin loss to ensure stable ranking between positive and negative pairs. The novelty of this work lies in the design of MedFuse-CLIP, a lightweight yet domain-tuned cross-modal retrieval framework that enhances fine-grained alignment between chest X-ray images and radiology reports.

In [4], Contrastive Visual Representation Learning (ConVIRT) was introduced for aligning chest X-rays and radiology reports through contrastive objectives, reporting a 6–8% improvement in retrieval accuracy over standard CNN-based baselines on the CheXpert dataset. GLoRIA [5] is a hierarchical global-local alignment model that combines patch-level and sentence-level matching. Using CheXpert and MIMIC-CXR, GLoRIA achieved Recall@10 ≈ 49.9 % for image→text retrieval and improved diagnostic classification AUROC by 4% compared to ResNet-based cross-attention models. BioViL [6] is a vision-language pretraining framework combining CNN-based visual encoders with BERT-style biomedical text encoders. On the MIMIC-CXR dataset, BioViL reported approximately 0.87 AUROC for zero-shot disease classification and achieved Recall@5 greater than 75% for cross-modal retrieval, outperforming generic CLIP variants. MedCLIP [7] is a radiology-specific adaptation of CLIP, trained using a combination of paired and unpaired medical images and text with prompt-based learning. MedCLIP achieved AUROC values exceeding 0.88 for pneumonia and cardiomegaly detection across multiple datasets and reduced inter-institutional domain shift by approximately 3% compared to baseline CLIP models. In cross-modal retrieval tasks, MedCLIP demonstrated competitive Recall@10 values (≈ 80–88%), although convergence was slower when training data were limited. CheXzero [8] is a zero-shot chest X-ray classifier, fine-tuned from CLIP using only radiology text corpora, without image-level labels. Evaluations on CheXpert and NIH-CXR14 showed AUROC values ranging from 0.88 to 0.90 for the most common pathologies, matching or surpassing fully supervised baselines. RadFuse [9] is a hybrid multimodal transformer with adaptive instance-level contrastive learning. RadFuse reported Recall@10 of 78.2 % and AUROC of 0.847 on MIMIC-CXR, demonstrating improved generalization under small-sample regimes, although performance in fine-grained semantic retrieval remained an open challenge. In [10], an optimized fine-tuning strategy for CLIP-based models was presented, which preserved joint image-text embedding alignment during retrieval adaptation. Experimental results showed consistent gains of 6–9% in Recall@5 and Recall@10 across image-text retrieval benchmarks, while maintaining semantic consistency between visual and textual embeddings.

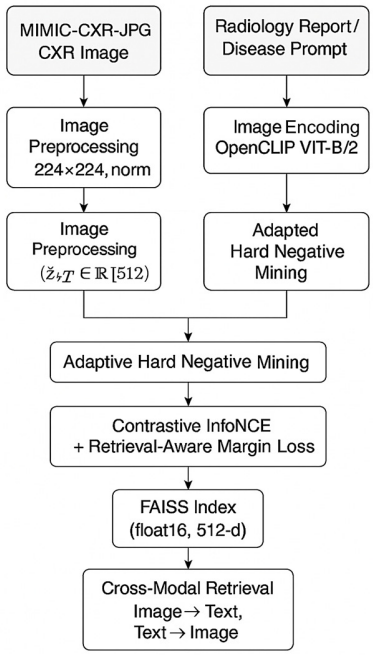

Fig. 1. Overview of the proposed approach.

II. MATERIALS AND METHODS

Figure 1 shows the end-to-end pipeline of the proposed MedFuse-CLIP model. Raw chest radiographs and their paired radiology reports are first preprocessed (resolution normalization and BPE tokenization, respectively). Both modalities are encoded using OpenCLIP ViT-B/32 into a shared 512-dimensional space [11]. Training is further regularized with a retrieval-aware margin loss to enforce rank stability. After training, all embeddings are stored in a FAISS index, enabling bidirectional image↔text retrieval on large-scale MIMIC-CXR-JPG using only 1–1.2 GB of memory [2].

A. Dataset Construction and Feature Representation

Experiments were conducted on the MIMIC-CXR-JPG dataset [12], a large-scale publicly available dataset containing approximately 370,000 chest radiographs paired with corresponding radiology reports. Each report includes structured sections (*Findings, Impression,* and *Comparison*), which provide rich textual supervision for image-text alignment. The JPEG version (377,110 images) was adopted to reduce preprocessing complexity compared to the original DICOM format, enabling training and retrieval on standard workstation hardware. Retrieval was formulated as a single-image↔single-report matching task. Each query image was paired with exactly one corresponding report. No study-level aggregation or multi-view fusion was employed. Only the *Findings* and *Impression* sections of the report were used as textual input.

1) Image Preprocessing

All radiographs were resized to a uniform resolution of 224×224 pixels and normalized using ImageNet statistics as:

$$I' = (I - \mu)/\sigma \quad (1)$$

where I denotes the original pixel values, I' the normalized image, μ = (0.485, 0.456, 0.406), and σ = (0.229, 0.224, 0.225).

2) Text Preprocessing

Each radiology report was tokenized with a Byte Pair Encoding (BPE) tokenizer [13] and padded or truncated to a fixed length of $L = 77$ tokens as given in:

$$T = \{t_1, t_2, \ldots, t_L\}, t_i \in \mathbb{Z}^+ \quad (2)$$

B. Model Architecture

The framework builds upon OpenCLIP ViT-B/32, consisting of two encoders trained in a contrastive manner [14].

- Image encoder:

$$f_\theta: \mathbb{R}^{H\times W\times C} \to \mathbb{R}^d, z_I = f_\theta(I') \in \mathbb{R}^d \quad (3)$$

which extracts a $d = 512$-dimensional visual embedding.

- Text encoder:

$$g_\phi: \mathbb{Z}^L \to \mathbb{R}^d, z_T = g_\phi(T) \in \mathbb{R}^d \quad (4)$$

which transforms the report tokens or prompt templates into textual embeddings.

Both modalities are projected to the unit hypersphere:

$$\hat{z}_I = \frac{z_I}{\|z_I\|_2}, \quad \hat{z}_T = \frac{z_T}{\|z_T\|_2} \quad (5)$$

so that cross-modal similarity is computed via cosine similarity,

$$s_{ij} = \hat{z}_{I_i}^\top \hat{z}_{T_j} \quad (6)$$

with an embedding dimension $D_g = 256$.

C. Adaptive Hard-Negative Mining

For each positive image-report pair, hard negatives were selected from samples belonging to the same patient or study but describing different clinical findings. Up to K = 4 hard negatives per batch were selected based on semantic similarity in the text embedding space. Negatives were weighted proportionally to their similarity score to emphasize difficult contrasts while avoiding false-negative dominance. Adaptive-hard-negative mining is introduced to refine discrimination among visually and semantically similar findings [15]. Patient-level splitting ensures that no clinically identical cases appear across splits, reducing the risk of false-negative contamination. For each image-report pair (I_i, T_i), negatives are selected from the same patient or study that depict different conditions:

$$\mathcal{N}_i = \{T_j \mid \text{Patient}(T_j) = \text{Patient}(T_i), Y_j \neq Y_i\} \quad (7)$$

Each negative is weighted by its semantic similarity to the positive text, computed using a Clinical-BERT encoder $h(\cdot)$:

$$w_{ij} = \frac{\exp(\text{sim}(h(T_i), h(T_j)))}{\sum_{n\in\mathcal{N}_i} \exp(\text{sim}(h(T_i), h(T_n)))}, \; j \in \mathcal{N}_i \quad (8)$$

D. Contrastive Learning with Retrieval-Aware Margin Loss

The training objective combines the standard symmetric InfoNCE contrastive loss and a retrieval-aware margin loss.

1) InfoNCE Loss

For a batch of N image–text pairs,

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[\log\frac{e^{s_{ii}/\tau}}{\sum_{j=1}^{N} e^{s_{ij}/\tau}} + \log\frac{e^{s_{ii}/\tau}}{\sum_{j=1}^{N} e^{s_{ji}/\tau}}\right] \quad (9)$$

where τ is a learnable temperature parameter controlling contrastive sharpness.

2) Margin Loss

To ensure positive pairs outrank hard negatives by a safety margin m,

$$\mathcal{L}_{\text{margin}} = \frac{1}{N}\sum_{i=1}^{N}\max\left(0, m + s_{i,n} - s_{ii}\right) \quad (10)$$

where $s_{i,n} = \max_{j\in\mathcal{N}_i} s_{ij}$ is the hardest negative similarity and $m = 0.2$.

3) Combined Objective

$$\mathcal{L} = \mathcal{L}_{\text{InfoNCE}} + \lambda\,\mathcal{L}_{\text{margin}} \quad (11)$$

with balancing weight $\lambda = 0.5$.

E. Training Configuration and Retrieval System

The model was optimized using AdamW with a learning rate of 5×10^{-6}, weight decay of 0.01, batch size of 128, temperature parameter τ = 0.07, margin m = 0.2, and loss weight λ = 0.5. Training was performed for 15 epochs with a linear warm-up over the first 500 optimization steps to stabilize early-stage contrastive learning and prevent gradient instability. All experiments were executed on a single NVIDIA RTX 3060 GPU with 12 GB of VRAM, demonstrating that the proposed approach is computationally feasible on standard workstation hardware without requiring large-scale clusters. The margin value m = 0.2 and balancing weight λ = 0.5 were selected based on preliminary sensitivity analysis on the validation split. Smaller margins (m< 0.1) resulted in insufficient separation between positive and hard-negative pairs, while larger margins (m > 0.3) led to unstable optimization and slower convergence.

III. EXPERIMENTAL RESULTS

All experiments were conducted using a strict patient-level split to prevent data leakage. The patients were divided into training (70%), validation (10%), and test (20%) sets with no overlap across splits. The retrieval index was constructed exclusively from the test split, while queries were sampled independently from the same split, ensuring that no patient, study, or image appeared in more than one split. A positive match was defined as a single chest X-ray image paired with its corresponding radiology report, and no study-level or multi-view aggregation was applied as shown in Table I. Retrieval indexing and querying were performed exclusively on the held-out test strict patient-level split. All reported metrics represent the mean over three independent runs with different random seeds. Performance variability across runs was minimal, indicating stable optimization behavior.

As summarized in Table II, MedFuse-CLIP achieves R@10 of 83.7% for image→text and R@10 of 79.6% for text→image retrieval, with median ranks of six and eight, respectively. As shown in Table III, the system achieved an average AUROC of 0.866 across 14 thoracic diseases, including 0.88 (Pneumonia), 0.87 (Cardiomegaly), and 0.89 (Pleural Effusion). Zero-shot disease classification was performed without any task-specific fine-tuning using fixed textual prompts derived from disease names. For each pathology, a corresponding text prompt was encoded using the frozen text encoder. Classification scores were obtained by computing cosine similarity between image embeddings and disease-specific text embeddings.

TABLE I. DATASET SPLITS UNDER STRICT PATIENT-LEVEL SEPARATION

Split	Patients	Studies	Images	Reports
Training	~46,000	~155,000	~258,000	~155,000
Validation	~6,500	~22,000	~37,000	~22,000
Test	~13,000	~45,000	~82,000	~45,000

TABLE II. CROSS-MODAL RETRIEVAL ON MIMIC-CXR (FINE-TUNED RETRIEVAL, VIT-B/32 BACKBONE)

Task	R@1 (%)	R@5 (%)	R@10 (%)	Median rank
Image→Text	62.4	78.5	83.7	6
Text→Image	58.9	74.1	79.6	8

TABLE III. ZERO-SHOT DISEASE CLASSIFICATION AUROC ON SELECTED THORACIC PATHOLOGIES.

Disease	AUROC
Pneumonia	0.88
Cardiomegaly	0.87
Pleural effusion	0.89
Atelectasis	0.85
Average (14 labels)	0.866

A. Ablation Study

Table IV summarizes a comprehensive ablation of negative-sample strategies and margin-based training. Starting from a baseline trained with random negatives, adding same-patient/study hard negatives improved R@10 by nearly 4%. Finally, incorporating the retrieval-aware margin loss (m = 0.2, λ = 0.5) yielded the highest scores of 87.5 % for image→text and 83.4 % for text→image, with the AUROC of 0.879 demonstrating that both novelties contribute additively to retrieval robustness. To isolate the effect of negative sampling, Table V details a comparison of three strategies for negatives.

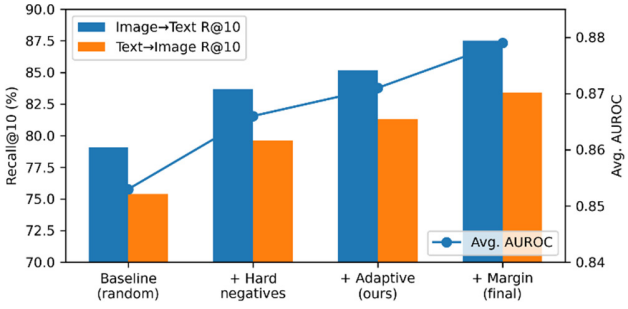

Fig. 2. Ablation analysis.

TABLE IV. EFFECT OF ADAPTIVE HARD-NEGATIVE MINING & MARGIN LOSS

Setting	Image→Text R@10 (%)	Text→Image R@10 (%)	Avg. AUROC
Baseline (InfoNCE+ random negatives)	79.1	75.4	0.853
+Standard hard negatives	83.7	79.6	0.866
+Adaptive semantic weighting (proposed)	85.2	81.3	0.871
+Margin loss (proposed)	87.5	83.4	0.879

TABLE V. NEGATIVE-SAMPLING STRATEGY COMPARISON

Negatives strategy	Image→Text R@10	Text→Image R@10	Avg. AUROC
Random (across dataset)	79.1	75.4	0.853
Same patient/study (hard)	83.7	79.6	0.866
Adaptive (proposed): semantic-weighted	85.2	81.3	0.871

B. Embedding Compression and Efficiency

To evaluate deployability on limited-memory systems, post-hoc PCA dimensionality reduction was applied to the learned 512-dimensional embeddings. Compressing embeddings to 256 dimensions reduced storage from 1.2 GB to 0.6 GB with only a loss of less than 2 % in R@10. As shown in Figure 3, retrieval performance remains nearly unchanged even after a 50% reduction in embedding dimensionality. The observed minimal accuracy degradation (<2 % at 256 dimensions) indicates that efficiency gains can be achieved without compromising retrieval reliability. FAISS indexing was implemented using an IVF-Flat index with cosine similarity. Embeddings were stored in float16 format, resulting in a memory footprint of approximately 1.2 GB for 512-dimensional embeddings and 0.6 GB after compression to 256 dimensions.

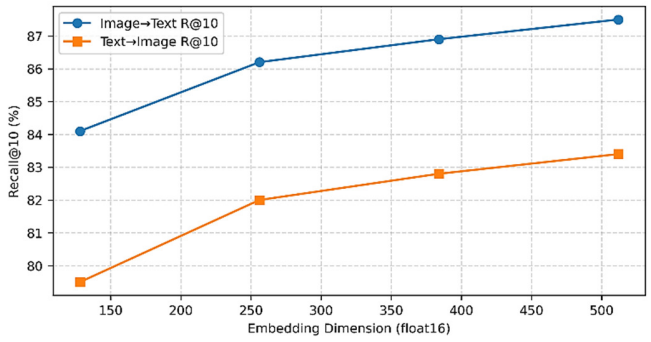

Fig. 3. Retrieval accuracy versus embedding dimensionality for MedFuse-CLIP using float16 compressed embeddings.

Table VI compares MedFuse-CLIP with representative radiology vision language models. Although datasets and evaluation splits differ, MedFuse-CLIP achieves retrieval accuracy comparable to or exceeding larger multi-GPU frameworks such as GLoRIA (ICCV 2021) and CXR-CLIP (2023).

TABLE VI. COMPARISON WITH EXISTING VISION–LANGUAGE MODELS.

Method	Backbone	Pretrain data	Retrieval (best reported)
GLoRIA [5]	ResNet-50	CheXpert pairs	49.9 (img→txt)
BioViL [6]	ViT/R50 + CXR-BERT	MIMIC-CXR	Strong ZS results reported
CXR-CLIP [2]	R50/Swin-T	MIMIC + CheXpert/CXR14	Outperforms BioViL/GLoRIA
RadFuse [9]	Transformer	MIMIC-CXR	78.2 / AUC 0.847
MedFuse-CLIP (proposed)	ViT-B/32	MIMIC-CXR-JPG	87.5 / 0.879 (single GPU)

It is important to note that the numerical comparisons in Table VI should not be interpreted as strict head-to-head benchmarks. Prior vision-language models such as GLoRIA, BioViL, and CXR-CLIP were evaluated under different experimental conditions, including varying datasets (e.g., CheXpert, NIH-CXR14, or mixed institutional corpora), split strategies, backbone architectures, and retrieval protocols (study-level or multi-view settings).

IV. CONCLUSION

This study introduced MedFuse-CLIP, a radiology-tuned cross-modal retrieval framework leveraging OpenCLIP ViT-B/32 for joint alignment of chest X-ray images and radiology reports. Through adaptive semantic hard-negative mining and a retrieval-aware margin loss, the model achieved robust visual-textual correspondence on the MIMIC-CXR dataset, obtaining Recall@10 of 87.5 % for image→text and 83.4 % for text→image, with a mean AUROC of 0.879 in zero-shot classification. Ablation results confirmed that adaptive negatives and margin constraints substantially improve retrieval precision, while 256-dimensional compressed embeddings preserved over 98% of accuracy at half the memory footprint. Compared with BioViL, GLoRIA, and CXR-CLIP, MedFuse-CLIP attains comparable or superior performance using a single-GPU setup, making it computationally efficient and reproducible. These results demonstrate that MedFuse-CLIP offers a scalable foundation for explainable radiology image-report retrieval and diagnostic reasoning in real-world clinical environments.

REFERENCES

[1] A. E. W. Johnson *et al.*, "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific Data*, vol. 6, no. 1, Dec. 2019, Art. no. 317, https://doi.org/10.1038/s41597-019-0322-0.

[2] K. You *et al.*, "CXR-CLIP: Toward Large Scale Chest X-ray Language-Image Pre-training," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, vol. 14221, H. Greenspan, A. Madabhushi, P. Mousavi, S. Salcudean, J. Duncan, T. Syeda-Mahmood, and R. Taylor, Eds. Springer Nature Switzerland, 2023, pp. 101–111.

[3] H. N. T. Al-Azzawi *et al.*, "Utilization of a Deep Convolutional Neural Network for the Binary Classification of Chest X-Ray Pneumonia," *Engineering, Technology & Applied Science Research*, vol. 15, no. 1, pp. 20471–20483, Feb. 2025, https://doi.org/10.48084/etasr.9788.

[4] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, "Contrastive Learning of Medical Visual Representations from Paired Images and Text," in *Proceedings of the 7th Machine Learning for Healthcare Conference*, Dec. 2022, pp. 2–25.

[5] S. C. Huang, L. Shen, M. P. Lungren, and S. Yeung, "GLoRIA: A Multimodal Global-Local Representation Learning Framework for

Label-efficient Medical Image Recognition," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 3922–3931, https://doi.org/10.1109/ICCV48922.2021.00391.

[6] B. Boecking *et al.*, "Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing," in *Computer Vision – ECCV 2022*, vol. 13696, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Springer Nature Switzerland, 2022, pp. 1–21.

[7] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "MedCLIP: Contrastive Learning from Unpaired Medical Images and Text," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 3876–3887, https://doi.org/10.18653/v1/2022.emnlp-main.256.

[8] J. Jang, D. Kyung, S. H. Kim, H. Lee, K. Bae, and E. Choi, "Significantly improving zero-shot X-ray pathology classification via fine-tuning pre-trained image-text encoders," *Scientific Reports*, vol. 14, no. 1, Oct. 2024, Art. no. 23199, https://doi.org/10.1038/s41598-024-73695-z.

[9] M. Jahanian, A. Karimi, N. O. Eraghi, and F. Zarafshan, "Multimodal transformers for joint diagnosis and retrieval from chest X-rays and radiology reports," *Informatics in Medicine Unlocked*, vol. 58, 2025, Art. no. 101672, https://doi.org/10.1016/j.imu.2025.101672.

[10] K. Schall, K. U. Barthel, N. Hezel, and K. Jung, "Optimizing CLIP Models for Image Retrieval with Maintained Joint-Embedding Alignment," in *Similarity Search and Applications*, vol. 15268, E. Chávez, B. Kimia, J. Lokoč, M. Patella, and J. Sedmidubsky, Eds. Springer Nature Switzerland, 2025, pp. 97–110.

[11] A. Nirala, A. Joshi, S. Sarkar, and C. Hegde, "Fast Certification of Vision-Language Models Using Incremental Randomized Smoothing," in *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, Apr. 2024, pp. 252–271, https://doi.org/10.1109/SaTML59370.2024.00019.

[12] A. Johnson *et al.*, "MIMIC-CXR-JPG - chest radiographs with structured labels." PhysioNet, https://doi.org/10.13026/JSN5-T979.

[13] F. Ashfaq, N. Jhanjhi, N. A. Khan, D. Javed, M. Masud, and M. Shorfuzzaman, "Enhancing ECG Report Generation With Domain-Specific Tokenization for Improved Medical NLP Accuracy," *IEEE Access*, vol. 13, pp. 85493–85506, 2025, https://doi.org/10.1109/ACCESS.2025.3567566.

[14] W. Dong, S. Shen, Y. Han, T. Tan, J. Wu, and H. Xu, "Generative Models in Medical Visual Question Answering: A Survey," *Applied Sciences*, vol. 15, no. 6, Mar. 2025, Art. no. 2983, https://doi.org/10.3390/app15062983.

[15] W. Huang, X. Hu, S. Abousamra, P. Prasanna, and C. Chen, "Hard Negative Sample Mining for Whole Slide Image Classification," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, vol. 15004, M. G. Linguraru, Q. Dou, A. Feragen, S. Giannarou, B. Glocker, K. Lekadir, and J. A. Schnabel, Eds. Springer Nature Switzerland, 2024, pp. 144–154.