Towards Comprehensive Benchmarking of Medical Vision Language Models

Dimple Khatri[1], Sanjan TP Gupta[2]
[1] University of Maryland Baltimore County, USA, [2] Indian Institute of Technology Madras, India

1. Background & Motivation

- Severe global radiologist shortage: India has 1 radiologist per 100,000 people, with even fewer in Tier-2/3 and rural regions.
- WHO reports complete absence of radiologists in parts of Africa, requiring tele-radiology for all interpretation.
- Even developed countries (UK, Japan, Singapore) face scan backlogs and radiologist fatigue, impacting diagnostic turnaround time.
- AI-assisted interpretation supports rapid triage, generates EHR-ready summaries, and improves diagnostic consistency.

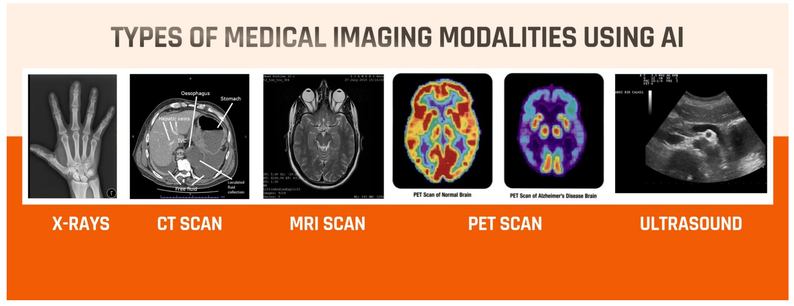

Figure-1: Examples of major medical imaging modalities, ranging from 2D grayscale X-rays to 3D CT/MRI volumes, PET functional scans, and real-time Ultrasound. These modalities differ in dimensionality, noise characteristics, and clinical use, motivating tailored AI approaches and the choice of X-rays as the starting point for Medical VLM benchmarking

2. Objective

- Establish a Phase-1 reproducible baseline for MedCLIP using the IU-CXR dataset, focusing on zero-shot image understanding.
- Evaluate how image preprocessing variants (RAW → MedCLIP preprocessing, CLAHE, Gaussian+CLAHE) affect interpretability and inference latency.
- Benchmark MedCLIP across **three core classification tasks**:
  - **View Identification:** Determine whether an X-ray is Frontal or Lateral.
  - **Multi-Label Disease Detection:** Identify all diseases present in an image, allowing multiple findings per X-ray.
  - **Single-Label Disease Classification:** Use a **stratified dataset with exactly one dominant pathology per image** to measure strict diagnostic accuracy.

Note: Multimodal retrieval and report summarization from the full MedVLM benchmarking framework will be introduced in Phase-2.

3. Datasets

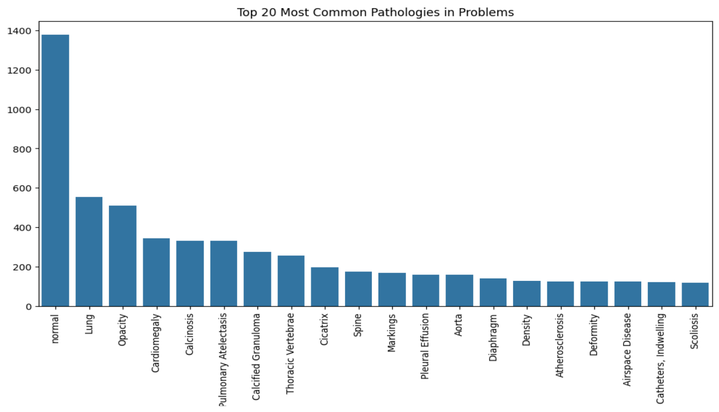

Figure-2: Top-20 pathology histogram

- **IU-CXR (Indiana University Chest X-ray Dataset):**

Public dataset of paired **chest X-ray images and radiology reports**.

Used for all Phase-1 experiments in this study.

- **Image Characteristics:**
  - 2D grayscale chest radiographs
  - Mixture of Frontal and Lateral views
  - Includes diverse pathologies and normal cases
  - Raw PNGs provided in varying resolutions
- **Label Sources Used:**
  - **View labels:** Derived from IU-CXR metadata
  - **Disease labels:** Extracted using **CheXpert-style keyword rules**
- **Custom Stratified Subset:**
  - Constructed to evaluate **single-label disease classification**
  - Each image assigned **one dominant pathology** for strict evaluation
  - Balances class distribution for fair comparison across diseases

Scan Here to connect on LinkedIn

Project Codebase Link Scan Here

****Open to research opportunities and collaborations**

4. Methodology

4.1 Model Architecture

- **Vision Encoder – Swin Transformer:**

MedCLIP uses a hierarchical Swin Transformer backbone that computes self-attention *within local windows*, enabling efficient scaling to high-resolution medical images.

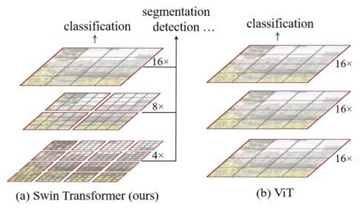

Figure 1. (a) The proposed Swin Transformer builds hierarchical feature maps by merging image patches (shown in gray) in deeper layers and has linear computation complexity to input image size due to computation of self-attention only within each local window (shown in red). It can thus serve as a general-purpose backbone for both image classification and dense recognition tasks. (b) In contrast, previous vision Transformers [20] produce feature maps of a single low resolution and have quadratic computation complexity to input image size due to computation of self-attention globally.

Figure-3: Swin vs ViT figure here; shows local window attention and hierarchical feature maps.

- **Text Encoder – BioClinicalBERT:**

A domain-tuned language model trained on clinical notes and radiology reports, providing medically grounded text embeddings.

- **Alternate Vision Backbone – ResNet-50:**

Used in comparative experiments to evaluate MedCLIPVisionModel performance differences.

- **Activation Function – GELU:**

Smoother and more stable than ReLU; standard in transformer-based architectures.

4.2 Zero-Shot Evaluation Pipeline

- Feed RAW X-ray images into MedCLIP.
- MedCLIP's internal image pipeline automatically performs:
  - Loading & resizing
  - Center cropping
  - Normalization to pretrained mean/std
  - Tensor conversion
- Images and prompts (e.g., "Frontal view", "Lateral view", "Cardiomegaly") are encoded into a shared contrastive space.
- Classification is computed through prompt similarity ranking or a 0.5 threshold for multi-label detection.

4.3 Preprocessing Variants (External to MedCLIP)

To study interpretability and latency, three input variants were evaluated **before** MedCLIP's internal preprocessing:

1. **RAW (labeled "Normalized" in figure; no external enhancement)**
2. **CLAHE-enhanced RAW** (contrast improvement)
3. **Gaussian Denoise + CLAHE** (noise reduction + contrast)

MedCLIP still applies its default preprocessing after these variants are passed into the model.

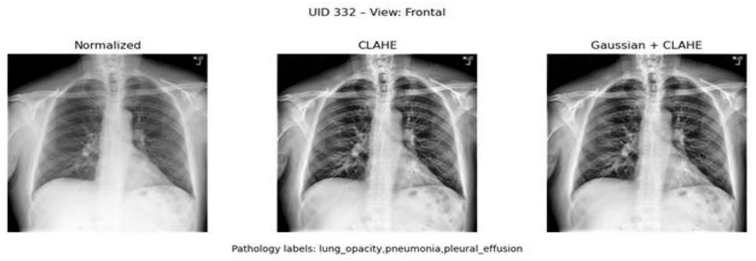

Figure-4: *Visual comparison of three preprocessing variants evaluated in Phase-1: RAW (labeled "Normalized"), CLAHE-enhanced, and Gaussian Denoise + CLAHE. These variants are used only for experimental analysis; MedCLIP's internal code still performs its default resizing and normalization during inference.*

Future Works:

1. Expand benchmarking to additional modalities and multi-center datasets to evaluate generalization beyond IU-CXR.
2. Introduce full multimodal tasks such as retrieval, report summarization, and structured RadGraph grounding.
3. Assess trustworthiness through calibration, rare-pathology performance, and robustness to perturbations.
4. Explore efficiency techniques (quantization, qLoRA, adapters) to balance accuracy with latency and memory use.
5. Extend benchmarking experiments to **CT, MRI, ophthalmology datasets**.

5. Experiments & Results

5.1 View Classification (Frontal vs Lateral)

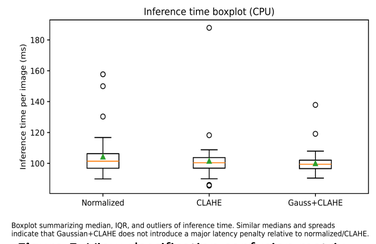

Figure-5: View classification confusion matrix (Frontal vs Lateral).

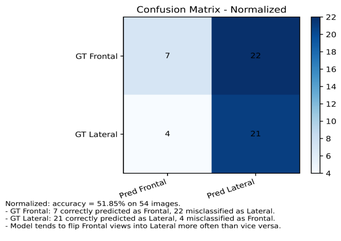

Figure-6: Inference latency boxplot for three preprocessing variants.

Key Result:

- MedCLIP achieves **~52% accuracy** on view identification.
- **No significant latency difference** across RAW, CLAHE, Gaussian+CLAHE variants.

Insights:

- Model **often flips Frontal → Lateral**, but rarely the reverse.
- Indicates weak global shape reasoning in zero-shot mode.
- Visual enhancement steps do **not** add computational burden on CPU.

5.2 Multi-Label Disease Detection

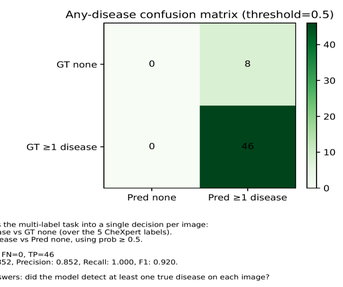

Figure-7: Any-disease confusion matrices (zero-shot threshold = 0.5)

Key Result:

Model shows **strong ranking ability** (AUROC ↑), but **low specificity** (many false positives).

Metrics:

Recall = 1.00 (no disease missed), **Precision = 0.69–0.85, F1 = 0.82–0.92**

5.3 Single-Label Disease Classification

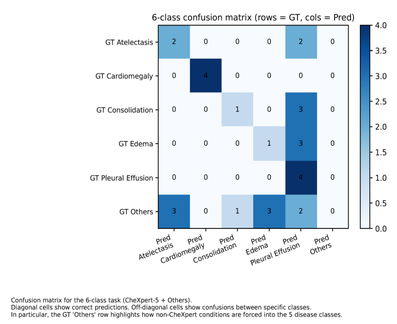

Figure-8: 6-class confusion matrix [Exp. 5.4]

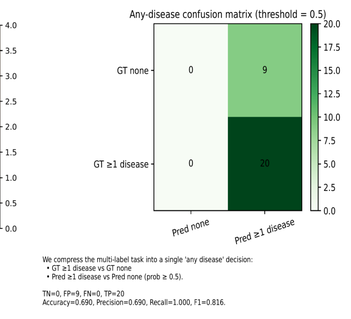

Figure-9: Strict single-label disease classification confusion matrix. [Exp. 5.3]

Key Result:

Model performs well on frequent diseases (e.g., Effusion) but poorly on rare long-tail categories.

Interpretation:

Highlights **label imbalance** and **zero-shot limitations** in fine-grained diagnosis.

5.4 Forced Single-Label Classification for 6-Class

When mapping multi-label disease predictions → 6 exclusive classes (CheXpert-5 + Others):

Accuracy **~41%**

Heavy confusion in **Others** class (not modeled by MedCLIP).

Demonstrates **task–model misalignment**, not model failure.

6. Conclusions

- MedCLIP is **strong at ranking** disease likelihoods (AUROC↑).
- Threshold-based classification is weak due to over-triggering (recall↑, specificity↓).
- View classification reveals challenge in **recognizing anatomical geometry**.
- Strict single-label experiment highlights **fine-grained limitations** of zero-shot MedCLIP.