MMed-RAG – Versatile Multimodal RAG System for Medical Vision-Language Models

Brij Kishor

Roll: 2411MC09

M.Tech (Mathematics and Computing)

Under the Supervision of

Dr. Jimson Mathew

Department of Computer Science & Engineering

Indian Institute of Technology Patna

Dataset Description

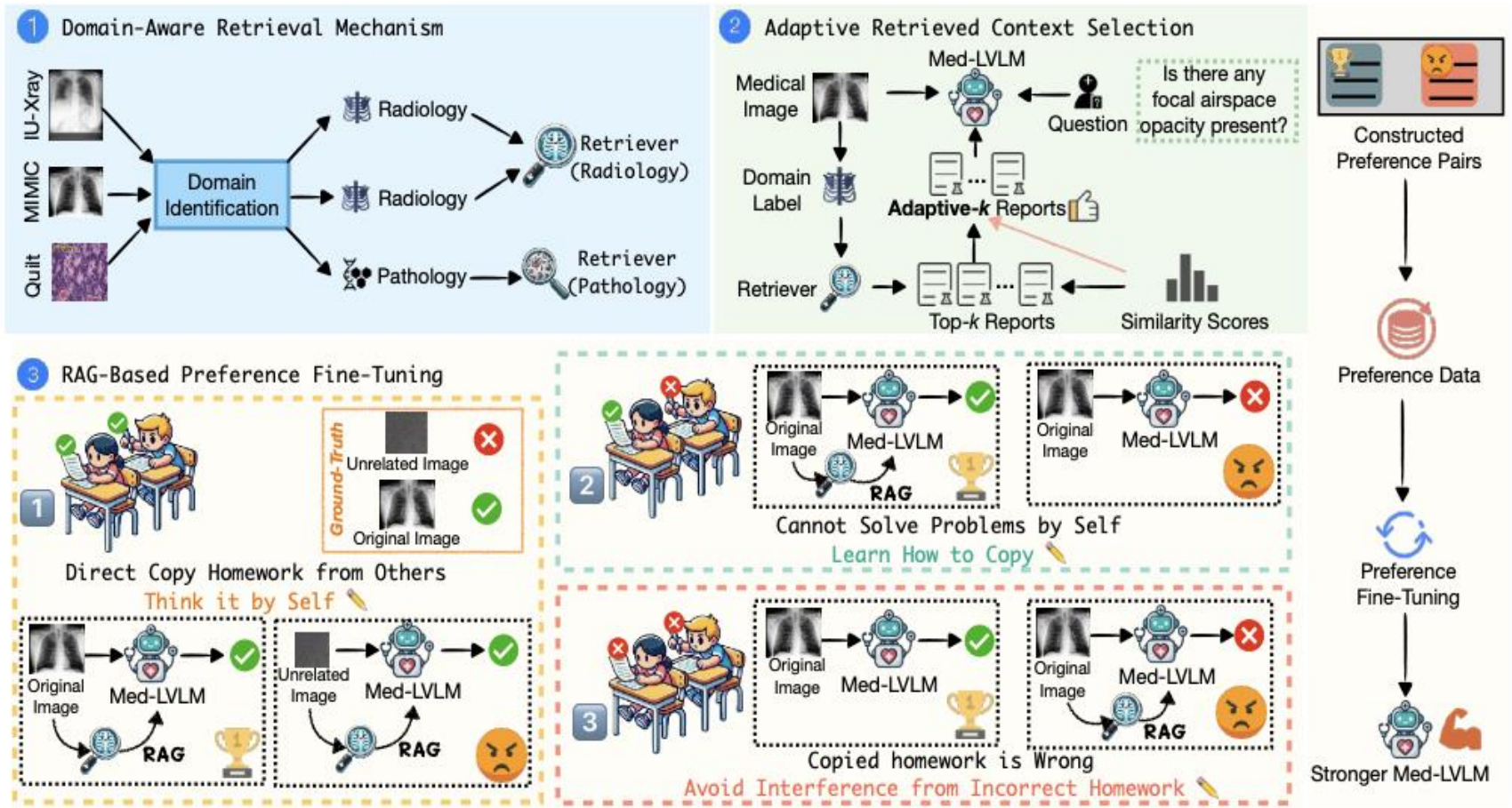
- Five diverse medical datasets were used:
- 1. MIMIC-CXR – chest X-rays and radiology reports.
- 2. IU-Xray – paired X-rays and short textual summaries.
- 3. FairVLMed (Ophthalmology) – retinal fundus images.
- 4. Quilt-1M – multimodal ophthalmic dataset.
- 5. PMC-Pathology – microscopic pathology images with annotations.
- Tasks include Visual Question Answering (VQA) and Report Generation.

- We have multiple Datasets class like **Generic Datasets, Report Generation Datasets, VQA Datasets classes**
- A Dataset class is like a smart container that stores information about where your data is ,It loads one sample at a time when asked.
- Preprocesses that sample (image + text),returns it in a format that the model can use

We have 2 type of files for each id

- 1.jpg files (slo_fundus_XXXXX.jpg)
- Purpose: Visual/image data
- Content: SLO (Scanning Laser Ophthalmoscopy) fundus images of the eye
- Use: The actual retinal images that would be used for visual analysis, deep learning models, or clinical review
- 2.npz files (data_XXXXX.npz)
- Purpose: Structured numerical/metadata storage
- Image embeddings or processed numerical representations of the fundus images
- Clinical features extracted from the images
- Metadata about the image (patient info, image quality metrics)
- Pre-computed features for machine learning models

Overview of MMed-RAG system



Bringing the Pretrained Model

- Instead of starting from scratch, the workflow loads a powerful pretrained model, such as an OpenCLIP ResNet50-CC12M variant already trained on millions of image–text pairs.
- This model already knows how to connect **images-text** because it has been trained on **12 million image–text pairs**.
- General CLIP is powerful, but **medical images (e.g., X-rays)** and **radiology/ Ophthalmology** reports are very different from everyday images.
- **Vision Encoder :-**A ResNet50-based feature extractor, Converts medical images into high-dimensional embeddings
- **Text Encoder :-**A transformer-based text model,Converts radiology report text into embeddings

CLIP-STYLE TRAINING PIPEINE: THE LEARNING LOOP

1. Model Creation

Build Image & Text Encoders,
Add Projection Heads (CLIP-style
(CLIP-style Architecture.

2. Data Loading & Setup

Prepare datasets (e.p. MIMIC-CX)
Use DistributedSamper for
(Multi-GPUR scheduler)

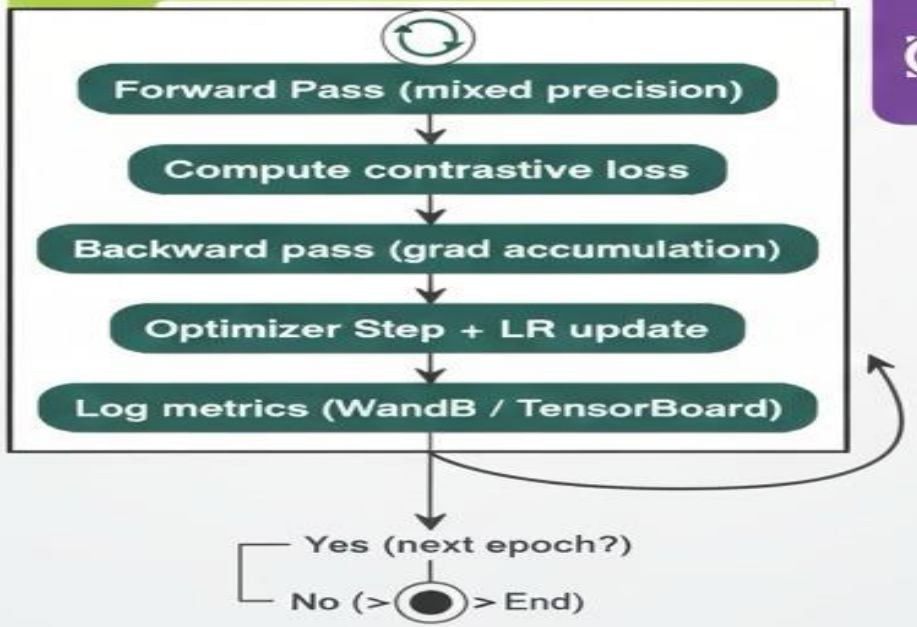
3. Optimizer & Loss Function

Create optimizer (AdamW).
Define Loss Configure LR
Scheduler

5. Evaluation

Extract features for full val set
Compute retrieval metrics
(R@1/5/10, mAP, Median Rank)

4. Training Loop (Per Epoch)



6. Checkpoint Saving

Save model + optimizer +
sptimizer + scheduler states
(checkpoint_epoch(n).pt)

Retrieval Process

- **Purpose:** Find most relevant reports for each test image using fine-tuned Medical CLIP.
- **Steps:**
  - Load fine-tuned CLIP model and checkpoint.
  - Encode all training reports (text encoder) and test images (image encoder).
  - Encode Test Images
  - Compute similarity scores between each test image and all reports.
  - Select top-K most similar reports for each test image.
- **Output:**
  - JSONL file with test image ID, ground truth report, and retrieved reference reports.
- **Key Points:**
  - Efficient, scalable batch retrieval.
  - Uses medical knowledge for accurate matching.
  - Enables knowledge-augmented training and evaluation in downstream tasks.

DPO

- **Purpose:** teach the language-vision policy to prefer image-grounded answers and to use retrieved text only when helpful, by directly comparing preferred vs dispreferred outputs (no reward model, no RL loops).
- **Inputs (per training run):**
  - Image queries (visual input).
  - Retrieval context(s) — top-k text snippets concatenated as extra context.
  - Candidate answers: a preferred (good) response and a dispreferred (bad) response for the same prompt+context.
  - Base policy (Med-LVLM) weights and tokenizer/config.
- **Outputs (what the process produces):**
  - Updated policy weights (periodic checkpoints).
  - Logs and metrics (preference accuracy, loss, KL drift, sample generations).
  - Optional serialized preference dataset used for reproducibility.

Cross-Modality Alignment

- Purpose: Ensure model's answers are grounded in the actual image and not blindly copied from retrieved text.
- Problem: Retrieved reports can dominate reasoning → confident but image-incorrect outputs.
- Approach: Create paired examples — (real image + question+retrieved context → correct answer) vs (counterfactual/noised image +same question+ same retrieved context → incorrect answer) — and train with a preference loss that ranks the real-image answer higher.
- Expected outcome: model uses retrieval only when consistent with visual evidence, reduces hallucinations, and improves image-grounded clinical accuracy.
- Key metric: preference accuracy (logp_preferred > logp_dispreferred)
- Large AI models sometimes ignore the image and just copy what sounds smart from the retrieved text.This training method forces the AI to pay attention to the image, not just guess from text.

Overall Alignment

- **Objective:** Teach the model to be image-grounded: use retrieved text only when it corroborates the image, ignore it when it conflicts.
- **Case 1 — Retrieval Helps**
  - When the image is ambiguous and retrieved reports are similar, retrieval provides useful evidence.
  - Desired behavior: model integrates retrieval + image and increases confidence in the correct interpretation.
- **Case 2 — Retrieval Misleads**
  - When retrieved reports are unrelated (e.g., report about pneumonia vs image showing a broken rib), retrieval is harmful.
  - Desired behavior: model prioritizes the image, ignores misleading retrieval, and reports the correct visual finding.
- **How training enforces balance**
  - Construct paired examples: (real image + retrieval → correct answer) vs (counterfactual/noisy image + same retrieval → wrong answer).
  - Use a preference loss (DPO): maximize log p(preferred) − log p(dispreferred) so the model favors image-grounded answers.
  - Include both retrieval-helpful and retrieval-misleading pairs so the model learns when to use or ignore retrieval.
- **Expected outcome**
  - Look at the image first; consult retrieval second.
  - Treat retrieved text only as supporting evidence. Always base the primary diagnosis on the visual content Reduced hallucinations; improved factual, image-grounded clinical outputs.

Results:-

Model	Rad	Opt	Pat
Med-Flamingo	27.42	22.50	29.11
MedVInT	33.17	29.40	25.33
RadFM	35.82	27.07	24.82
miniGPT-Med	36.66	25.28	23.16
MMed-RAG	**56.94**	**56.38**	**54.10**

Model	IU-Xray		FairVLMed	
	VQA	RG	VQA	RG
LLaVA-Med-1.5	68.99	10.04	66.63	13.41
+DR	77.12	13.23	72.69	15.89
+RCS	79.56	17.92	75.74	17.22
+RAG-PT (Ours)	**85.80**	**29.80**	**87.18**	**20.42**

- Ablation studies show:
- - Domain Retrieval adds +17–18% accuracy.
- - Adaptive Context Selection adds +6–19%.
- - Preference Fine-Tuning adds +16–37%.
- Theoretical validation ensures increased image dependence and better retrieval alignment.
- Attention visualization confirmed stronger grounding on image regions post fine-tuning.

Future Work and Enhancements

- Integrate real-time factual verification using medical knowledge graphs (UMLS, RadGraph).
- Implement dynamic RAG activation – use retrieval only when model uncertainty is high.
- Introduce attention-based explainability dashboards for clinical transparency.
- Extend MMed-RAG to multilingual medical datasets for global adaptability
- Explore integration with continual learning to adapt to evolving medical knowledge.

Thanks for your attention!!