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

Brij Kishor

Roll: 2411MC09

M.Tech (Mathematics and Computing)

MTP Presentation Under the Supervision of

Dr. Jimson Mathew

Department of Computer Science & Engineering

Indian Institute of Technology Patna

Index:-

- Problem statement & Literature Survey
- Dataset Description
- Bringing the Pretrained Model
- Architecture
- CLIP style Training Pipeline
- Retrieval Process After Training
- Cross-Modality Alignment
- Overall Alignment
- RAG based preference-fine tuning
- LoRA Fine-tuning in Medical Multimodal DPO Training
- Experimental Results
- Conclusion
- Future Work and Enhancements

Problem statement & Literature Survey

- Problem Statement :-Develop a domain-aware, retrieval-augmented framework that enhances medical LVLMs by reducing hallucinations, improving factual consistency, and enabling generalization across diverse medical imaging domains.
- Literature Survey:-

Earlier Med-LVLMs like Med-Flamingo, RadFM, LLaVA-Med show strong visual-language alignment but **high hallucination rates.**

Prior RAG systems (e.g., RULE, MedDR, FactMM-RAG) improved factuality but **lacked domain specialization or alignment control**.

RAG in general AI improved factual grounding but medical RAGs struggled with **domain mismatch and over-reliance on retrieval**.

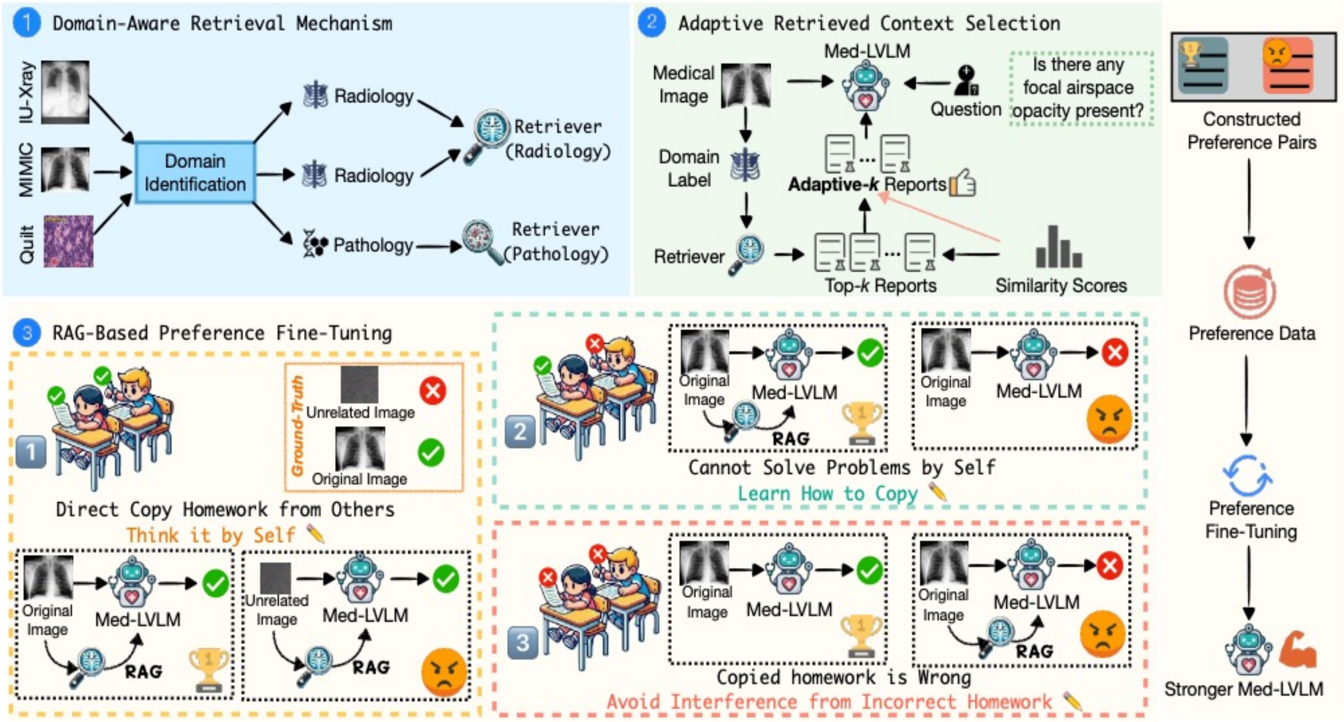
Dataset Description:-

- 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**

Bringing the Pretrained Model:-

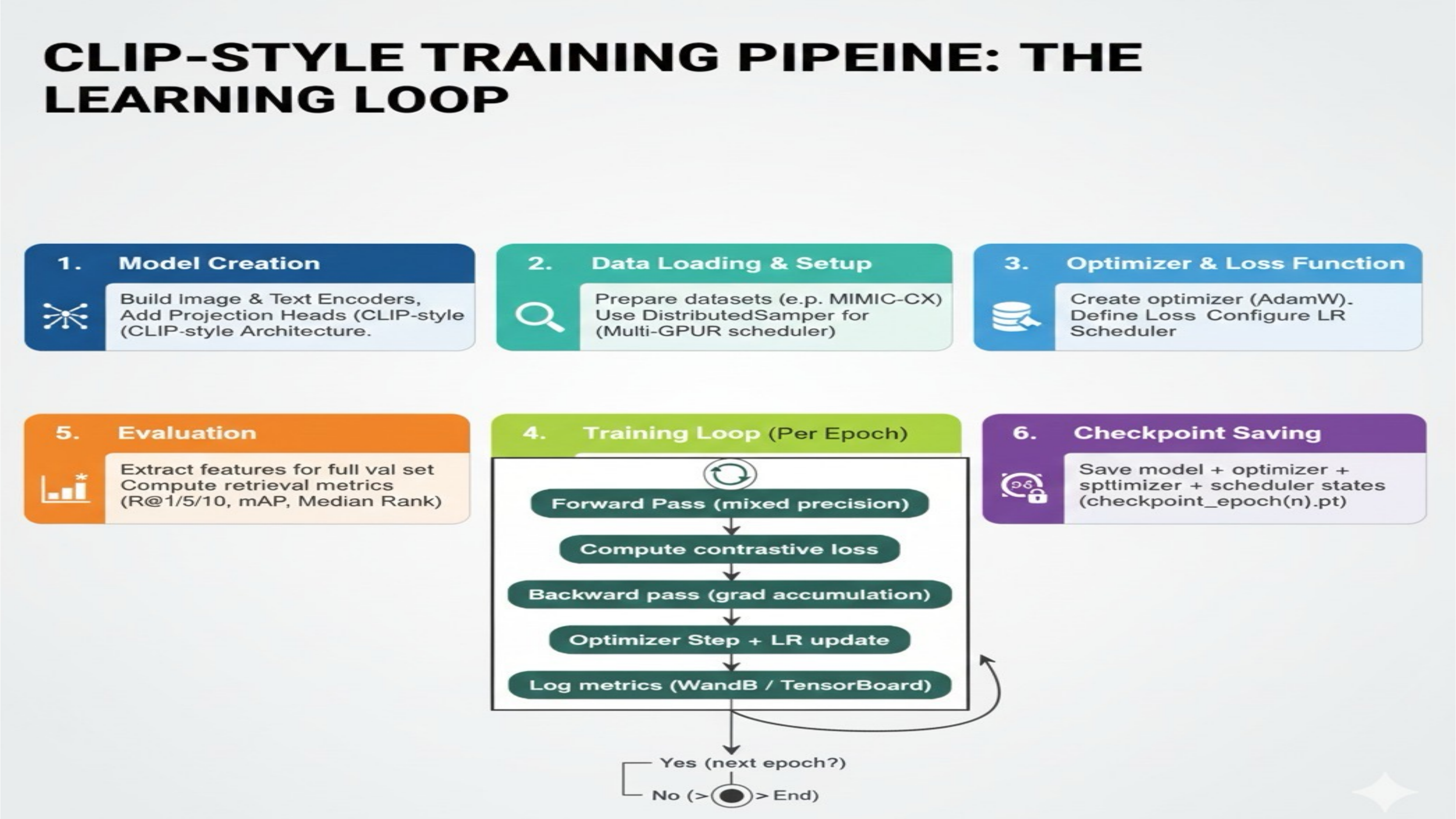
- Instead of starting from scratch, the workflow loads a powerful pretrained model, such as an OpenCLIP ResNet50-CC12M variant already trained on millions of image-text pairs.
- This model already knows how to connect **images-text** because it has been trained on **12 million image-text pairs**.
- General 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 medical report text into embeddings

Architecture:-



Steps:-

- **Domain-Aware Retrieval :** MMed-RAG first classifies the input image into a domain (e.g. radiology vs pathology) using a vision-language model (BiomedCLIP) that was fine-tuned on a small labeled set. Once the domain is known, the image is passed to a retriever trained in that domain. then it returns the top-k most similar reports. Those become the *retrieved contexts* for that image
- **Adaptive Context Selection:** Second, instead of always taking a fixed number of retrieved documents, MMed-RAG dynamically decides how many to use based on how **similar** they are to the image. In many retrieval tasks, the relevance of documents drops sharply after a point. For example, the 5th most relevant report might be almost as good as the 1st, but by the 10th report the similarity score suddenly drops.



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)

- Forward Pass (mixed precision)
- Compute contrastive loss
- Backward pass (grad accumulation)
- Optimizer Step + LR update
- Log metrics (WandB / TensorBoard)

Yes (next epoch?)

No (> ● > End)

6. Checkpoint Saving

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

Dataloader input-

- Raw JSON data (7000 entries)
- data: [
- {
- "id": "data_0",
- "filename": "data_0.npz",
- "image_path": "Left-Fundus/1.png",
- "gpt4_summary": "The image shows a retinal fundus photograph of the left eye...",
- "use": "training"
- },],
- Processed image-report pairs (7000 tuples)
- image_report_pairs: [
- (
- "/…/1.png",
- "The image shows a retinal fundus photograph of the left eye..."
-),
- # 3. Image IDs (7000 strings)
- image_ids: [
- "data_0",
- "data_1",
- "data_2",…]

During iteration DataLoader does

- for batch in dataloader:
  - Picks 4 random indices: [234, 891, 45, 1023]
  - Calls dataset[234], dataset[891], dataset[45], dataset[1023]
  - Each call triggers dataset.__getitem__() which:
  - Loads image from disk
  - Applies transforms
  - Tokenizes text
  - Returns (img_tensor, text_tensor, path)
  - Collates 4 samples into batched tensors
  - Returns batched data
- Batched into tensors are ready for model

Retrieval Process After Training:-

- **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 and images .
  - 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.
- **DPO:** 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).
- Input-> image + question +retrieved context
- Output->Updated policy weights (periodic checkpoints).

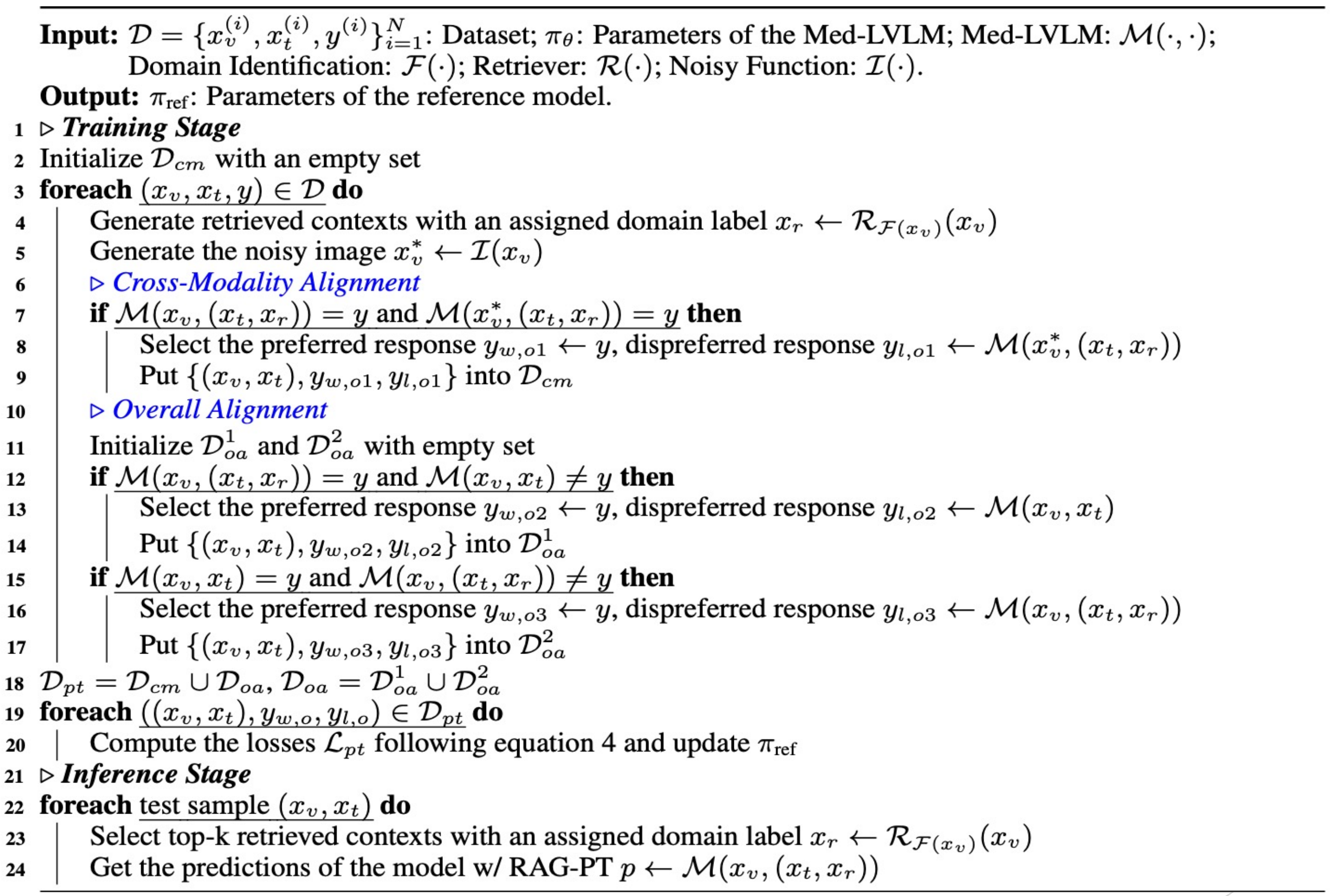
Cross-Modality Alignment :-

- Purpose: Ensure model’s answers are grounded in the actual image and not blindly copied from retrieved text.
- 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:** use retrieved text only when it helps , 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.
- **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.

RAG based preference-fine tuning:-



```
Input: $\mathcal{D} = \{x_v^{(i)}, x_t^{(i)}, y^{(i)}\}_{i=1}^{N}$: Dataset; $\pi_\theta$: Parameters of the Med-LVLM; Med-LVLM: $\mathcal{M}(\cdot,\cdot)$;
       Domain Identification: $\mathcal{F}(\cdot)$; Retriever: $\mathcal{R}(\cdot)$; Noisy Function: $\mathcal{I}(\cdot)$.
Output: $\pi_{\text{ref}}$: Parameters of the reference model.
▷ Training Stage
Initialize $\mathcal{D}_{cm}$ with an empty set
foreach $(x_v, x_t, y) \in \mathcal{D}$ do
    Generate retrieved contexts with an assigned domain label $x_r \leftarrow \mathcal{R}_{\mathcal{F}(x_v)}(x_v)$
    Generate the noisy image $x_v^* \leftarrow \mathcal{I}(x_v)$
    ▷ Cross-Modality Alignment
    if $\mathcal{M}(x_v, (x_t, x_r)) = y$ and $\mathcal{M}(x_v^*, (x_t, x_r)) = y$ then
        Select the preferred response $y_{w,o1} \leftarrow y$, dispreferred response $y_{l,o1} \leftarrow \mathcal{M}(x_v^*, (x_t, x_r))$
        Put $\{(x_v, x_t), y_{w,o1}, y_{l,o1}\}$ into $\mathcal{D}_{cm}$
    ▷ Overall Alignment
    Initialize $\mathcal{D}_{oa}^1$ and $\mathcal{D}_{oa}^2$ with empty set
    if $\mathcal{M}(x_v, (x_t, x_r)) = y$ and $\mathcal{M}(x_v, x_t) \neq y$ then
        Select the preferred response $y_{w,o2} \leftarrow y$, dispreferred response $y_{l,o2} \leftarrow \mathcal{M}(x_v, x_t)$
        Put $\{(x_v, x_t), y_{w,o2}, y_{l,o2}\}$ into $\mathcal{D}_{oa}^1$
    if $\mathcal{M}(x_v, x_t) = y$ and $\mathcal{M}(x_v, (x_t, x_r)) \neq y$ then
        Select the preferred response $y_{w,o3} \leftarrow y$, dispreferred response $y_{l,o3} \leftarrow \mathcal{M}(x_v, (x_t, x_r))$
        Put $\{(x_v, x_t), y_{w,o3}, y_{l,o3}\}$ into $\mathcal{D}_{oa}^2$
$\mathcal{D}_{pt} = \mathcal{D}_{cm} \cup \mathcal{D}_{oa}$, $\mathcal{D}_{oa} = \mathcal{D}_{oa}^1 \cup \mathcal{D}_{oa}^2$
foreach $((x_v, x_t), y_{w,o}, y_{l,o}) \in \mathcal{D}_{pt}$ do
    Compute the losses $\mathcal{L}_{pt}$ following equation 4 and update $\pi_{\text{ref}}$
▷ Inference Stage
foreach test sample $(x_v, x_t)$ do
    Select top-k retrieved contexts with an assigned domain label $x_r \leftarrow \mathcal{R}_{\mathcal{F}(x_v)}(x_v)$
    Get the predictions of the model w/ RAG-PT $p \leftarrow \mathcal{M}(x_v, (x_t, x_r))$
```

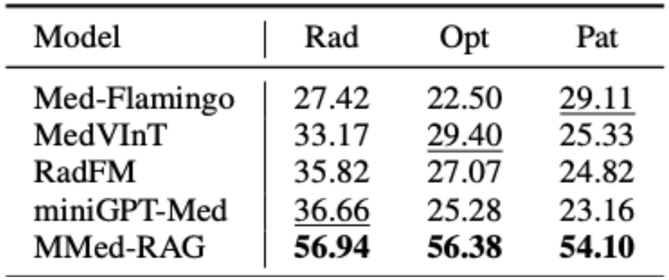
LoRA Fine-tuning in Medical Multimodal DPO Training

- **Problem:** Fine-tuning large models (7B+ parameters) is expensive
- **Solution:** Only train small "adapter" matrices instead of full model
- **Key Insight:** Model updates have low intrinsic dimensionality
- Wnew = Woriginal + ΔW

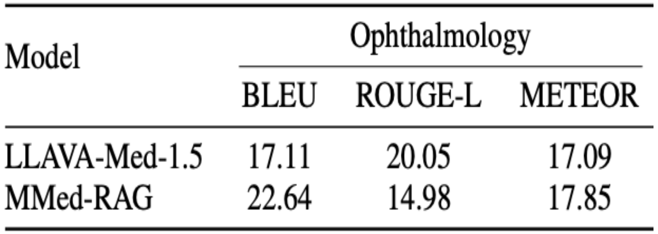
ΔW = A × B # where A: d×r, B: r×d, r << d

Metric	Full Fine-tuning	LoRA (r=64)	Savings
Trainable Parameters	7B	~16M	**99.8%** reduction
GPU Memory	28GB	12GB	**57%** reduction
Training Speed	1x	1.5x	**50%** faster
Storage	28GB	64MB	**99.9%** reduction

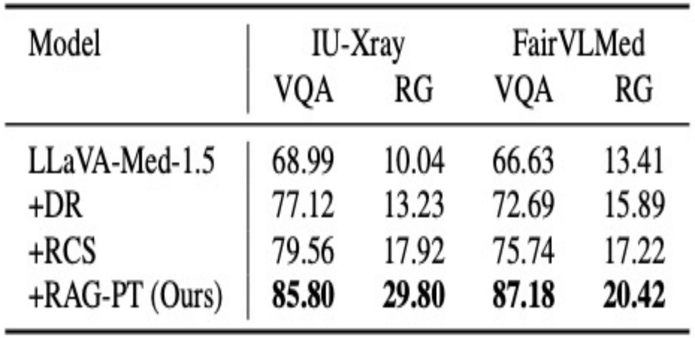
Experimental Results:-



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	Ophthalmology		
	BLEU	ROUGE-L	METEOR
LLAVA-Med-1.5	17.11	20.05	17.09
MMed-RAG	22.64	14.98	17.85



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**

Conclusion :-

- MMed-RAG effectively addresses hallucination issues in medical multimodal models.
- Combines retrieval grounding with fine-tuning alignment for factual and explainable outputs.
- Provides a scalable framework for cross-domain medical reasoning.
- Lays foundation for future factuality-aware, interpretable, and trustworthy medical AI systems.

Future Work and Enhancements:-

- **Medical Knowledge Graph Integration**

Modify: Data loading to include KG annotations

First integration of medical KGs with multimodal DPO

Real-time fact verification during generation that will help to reduce medical hallucinations.

If a statement conflicts with the KG, we penalize that continuation and push the model towards KG-consistent answers, further reinforced using DPO with KG-based feedback.

- **Multilingual Medical Dataset Extension**

Current MMed-RAG is English-only. Extend to multilingual for global healthcare.

Modify: Data preprocessing to handle multiple languages

First multilingual medical multimodal DPO system

Preserves medical accuracy across language translations

Enables global deployment of medical AI

Future Work and Enhancements:-

- **Few-Shot Adaptation for Rare Medical Conditions**

Problem: Rare diseases have extremely limited data (**only 3-5 patient examples**)

Impact: Affects **400M people**, **7,000+ diseases**

Goal: Enable AI to learn and adapt from very few samples

Outcome: Faster, accurate diagnosis — reducing time from **years to minutes**

Thanks for your attention!!