Multimodal Retrieval System Report

1. Introduction

This project implements a multimodal retrieval engine designed for clinical chest X-ray data. The system integrates image and text modalities, enabling flexible retrieval across text→text, text→image, image→text, image→image, and multimodal queries. The workflow emphasizes reproducibility, efficient preprocessing, and CPU-friendly inference, making it suitable for deployment in constrained environments while maintaining clinical relevance.

---

2. Methods

2.1 Dataset Preparation

A random sample of 500 records was extracted from the MIMIC-CXR parquet file. Each record contained raw image bytes and associated metadata (*findings* and *impressions*). Images were decoded into `.png` format and stored in a dedicated directory, while metadata was indexed in a CSV file. This ensured traceability between images and their diagnostic context.

2.2 Text Chunking

Findings and impressions were segmented into chunks of up to 128 tokens using a custom sentence-based splitting function. This produced granular text units suitable for embedding models, serialized into a JSONL file (`multimodal_chunks.jsonl`) alongside image paths and section labels.

2.3 Embedding Generation

Text Models Evaluated:

- **MiniLM-L6-v2**: Groups lexically similar text (0.042s latency, 384-dim, 82.8MB memory)
- **Multi-QA MiniLM**: Groups semantically similar text (0.046s, 384-dim, 43.1MB) - **Selected**
- **Bio_ClinicalBERT**: Groups clinically related concepts (0.146s, 768-dim, 27.3MB)

Multi-QA MiniLM was selected for its semantic understanding and efficiency. Bio_ClinicalBERT showed superior clinical nuance (linking devices, interventions) but 3x slower with 2x embedding size.

Vision Models Evaluated:

- **CLIP ViT-B/32**: 0.09s latency, 512-dim, 364.5MB - **Selected**
- **CLIP ViT-L/14**: 1.72s latency, 768-dim, 1176.5MB

ViT-B/32 selected for speed-quality balance.

- **Image embeddings**: Extracted using the CLIP model (`openai/clip-vit-base-patch32`). Images were processed in batches, converted to RGB, and encoded into dense vectors.
- **Text embeddings**: Generated with a sentence-transformer (`multi-qa-MiniLM-L6-cos-v1`). Text chunks were tokenized, passed through the transformer, and pooled via mean aggregation.
  Both embeddings were saved as NumPy arrays for downstream indexing.

2.4 Index Construction

Embeddings were normalized to unit length and indexed using **FAISS** with inner-product similarity. Separate indices were built for text and image embeddings, while a fused multimodal index was created by concatenating aligned text and image vectors. This enabled unified retrieval across modalities.

2.5 Retrieval Application

A **Streamlit UI** was developed to expose the retrieval modes:

- **Text→Text**: semantic search over textual chunks
- **Text→Image**: CLIP-based text→image retrieval
- **Image→Text**: image query mapped to textual evidence
- **Image→Image**: similarity search among images
- **Multimodal→Multimodal**: joint query combining text and image vectors

2.6 Answer Generation

Retrieved evidence was passed into a **Flan-T5** model (CPU-friendly) via structured prompts. The model synthesized cohesive clinical summaries or image similarity descriptions. If evidence was insufficient, the system returned “Unanswerable” to preserve reliability. Evidence lines were optionally displayed in an expandable section for transparency.

---

3. Results

- **Dataset outputs**: 500 chest X-ray images, metadata CSV, and chunked JSONL file.
- **Embeddings**: Dense vectors for both text and images, stored as `.npy` arrays.
- **Indices**: FAISS indices (`text.index`, `vision.index`, `multimodal.index`) enabling efficient similarity search.

- **UI demo**: Interactive retrieval across five modes, with generated answers and expandable evidence display using streamlit.

---

4. Discussion

The pipeline demonstrates a reproducible workflow for multimodal retrieval in clinical imaging. Strengths include modular preprocessing, batch-friendly embedding generation, and CPU-optimized answer synthesis. Limitations include reliance on fixed chunk sizes and potential misalignment between text and image samples. Future work could explore adaptive chunking, larger multimodal fusion models, and integration with evaluation metrics for clinical QA.

---

5. Conclusion

This system provides a complete end-to-end framework for multimodal retrieval, combining dataset preparation, embedding generation, FAISS indexing, and interactive querying. By aligning text and image modalities, it enables richer exploration of clinical datasets and lays the groundwork for explainable, human-centered AI in healthcare.