M3Retrieve: Benchmarking Multimodal Retrieval for Medicine

Arkadeep Acharya[1*†] **Akash Ghosh**[1*] **Pradeepika Verma**[1]
Kitsuchart Pasupa[2] **Sriparna Saha**[1] **Priti Singh**[1]

[1]Indian Institute of Technology Patna, India
[2]King Mongkut's Institute of Technology Ladkrabang, Thailand

Abstract

With the increasing use of Retrieval-Augmented Generation (RAG), strong retrieval models have become more important than ever. In healthcare, multimodal retrieval models that combine information from both text and images offer major advantages for many downstream tasks such as question answering, cross-modal retrieval, and multimodal summarization, since medical data often includes both formats. However, there is currently no standard benchmark to evaluate how well these models perform in medical settings. To address this gap, we introduce ***M3Retrieve***, a Multimodal Medical Retrieval Benchmark. ***M3Retrieve***, spans 5 domains,16 medical fields, and 4 distinct tasks, with over 1.2 Million text documents and 164K multimodal queries, all collected under approved licenses. We evaluate leading multimodal retrieval models on this benchmark to explore the challenges specific to different medical specialities and to understand their impact on retrieval performance. By releasing ***M3Retrieve***, we aim to enable systematic evaluation, foster model innovation, and accelerate research toward building more capable and reliable multimodal retrieval systems for medical applications. The dataset and the baselines code are available in this github page https://github.com/AkashGhosh/M3Retrieve.

1 Introduction

Retrieval models play a crucial role in efficiently accessing and utilizing the vast amounts of information available today. These models facilitate the quick and accurate extraction of relevant data, essential for informed decision-making in various downstream applications across numerous domains. With the emergence of Retrieval-Augmented Generation (RAG) (Lewis et al., 2021), the importance of high-quality retrieval systems has grown exponentially, particularly in knowledge-intensive domains.

In recent years, advancements in deep learning have facilitated the development of multimodal retrieval models that process and generate embeddings from both textual and visual data. This capability is particularly significant in the medical domain, where images such as X-rays, MRIs, and histopathological slides provide critical context alongside textual descriptions, a fact that has already been highlighted in works like LLaVa-Med (Li et al., 2023) and MedSumm (Ghosh et al., 2024b). The importance of both image and text embedding for effective knowledge extraction for the Medical domain has been highlighted in existing works like (Ghosh et al., 2024a,d). The performance of multimodal retrievers is vital for downstream tasks such as multimodal information extraction (Sun et al., 2024), question answering (Luo et al., 2023), and cross-modal retrieval (Wang et al., 2025), as it directly impacts the accuracy of generated content. This becomes especially critical in safety-sensitive domains like healthcare, where trust and reliability are paramount. However, despite these developments, there is currently no standardized benchmark to evaluate the performance of multimodal retrieval models in medical applications.

Research Gap: Though efforts have been made to develop extensive retrieval benchmarks, including the BEIR Benchmark (Thakur et al., 2021) and the M-BEIR Benchmark (Wei et al., 2023) for the evaluation of text-only retrievals and multimodal retrievals, respectively, their expansion into domain-specific tasks, such as medical retrieval, remains an open challenge. Besides being an important domain of study for NLP and machine learning applications in general, the importance of benchmarking in this domain cannot be overstated. The medical field is highly complex, and access to precise and

[*] Equal contribution. Contact author: acharyarka17@gmail.com, akash_2321cs19@iitp.ac.in

[†] Currently working at IBM Research, Bangalore, India.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 15275–15288
November 4-9, 2025 ©2025 Association for Computational Linguistics

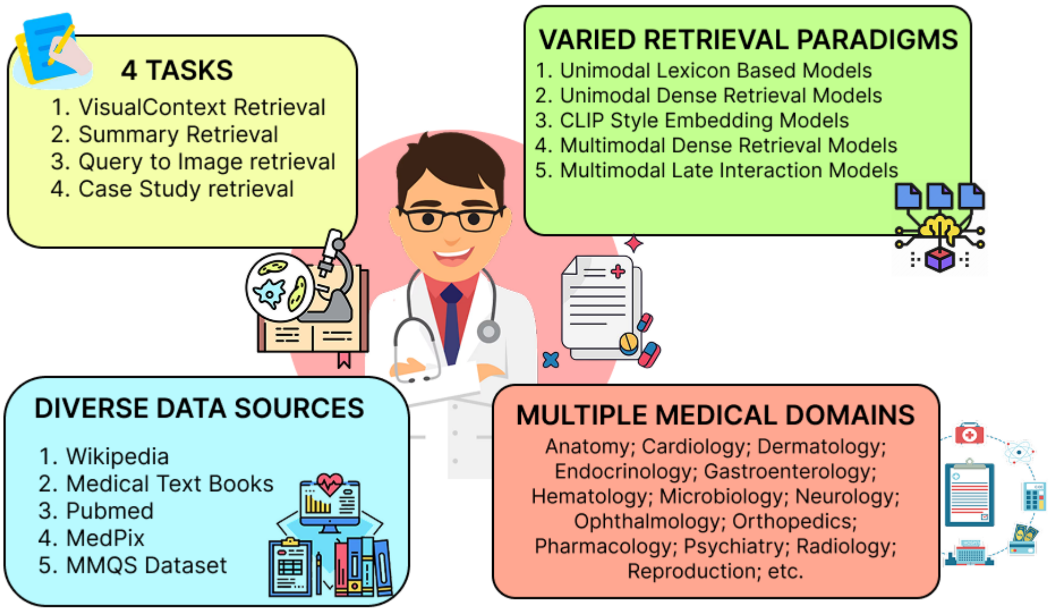

4 TASKS
1. VisualContext Retrieval
2. Summary Retrieval
3. Query to Image retrieval
4. Case Study retrieval
VARIED RETRIEVAL PARADIGMS
1. Unimodal Lexicon Based Models
2. Unimodal Dense Retrieval Models
3. CLIP Style Embedding Models
4. Multimodal Dense Retrieval Models
5. Multimodal Late Interaction Models
DIVERSE DATA SOURCES
1. Wikipedia
2. Medical Text Books
3. Pubmed
4. MedPix
5. MMQS Dataset
MULTIPLE MEDICAL DOMAINS
Anatomy; Cardiology; Dermatology; Endocrinology; Gastroenterology; Hematology; Microbiology; Neurology; Ophthalmology; Orthopedics; Pharmacology; Psychiatry; Radiology; Reproduction; etc.

Figure 1: *M3Retreive* is a multimodal medical retrieval benchmark comprising samples from **four** different tasks across **multiple healthcare subdomains** obtained from a variety of open-sourced data sources resulting in total dataset size of about 800K query-corpus pairs. It encompasses testing of retrieval models across varied retrieval paradigms

relevant information can significantly influence patient care and medical research. An intricate field like Medicine poses a unique set of challenges, such as: a) **Complex Medical Terminologies:** The medical field uses complex, variable terminology that often requires plain-language explanations for clarity; for instance, *Acute Hemorrhagic Leukoencephalitis*—a severe brain inflammation—may be described as causing *a sudden, severe headache and episodes of confusion*[1], highlighting the need for retrieval systems to accurately interpret such terms. b) **Multiple niche specialities:** Medicine is divided into many specialized disciplines, each requiring tailored methods to address specific patient needs[2][3]; as hospitals are organized by specialty, retrieval systems must be finely evaluated to assess their generalization across diverse medical domains. c) **Complex Image-Text Relationship:** Medical images can appear similar yet represent different conditions when combined with patient history; for example, *Viral Exanthems* and *Drug Eruptions* are hard to distinguish without detailed context[4], requiring multimodal retrieval systems to jointly encode and interpret image-text data accurately.

Existing medical datasets in the BEIR Benchmark focus on a **single (textual) modality** and lack the necessary scale and diversity required for fine-grained assessment of retrieval performance across **multiple medical disciplines**. Medical text-image pair datasets have not been explicitly covered in any of the datasets included in the M-BEIR Benchmark, thus creating a pressing need for a comprehensive benchmark tailored to the medical domain that evaluates retrieval models on real-world, multimodal data. Table 1 provides a comprehensive overview of existing multimodal medical datasets for retrieval, highlighting how ***M3Retrieve***distinguishes itself in terms of task complexity and the broad range of domains it covers.

Present Work: We introduce ***M3Retrieve***, a Multimodal Medical Retrieval Benchmark designed to bridge the gap in medical information retrieval. By integrating both textual and visual data, **M3Retrieve** enables a more realistic evaluation of retrieval models in complex multimodal medical contexts. The key contributions of this

[1] https://www.medicalnewstoday.com/articles/acute-hemorrhagic-leukoencephalopathy

[2] https://en.wikipedia.org/wiki/Medical_specialty

[3] https://www.abms.org/wp-content/uploads/2021/12/ABMS-Guide-to-Medical-Specialties-2022.pdf

[4] https://www.slideshare.net/slideshow/viral-exanthemsmodule/38060053

Table 1: Comparison of Benchmark Retrieval Datasets in the Medical Domain

Benchmark	**Data Points**	**Task**	**Medical Domains**	**Modalities**	**Open?**
NFCorpus (Boteva et al., 2016)	3,244 queries, 9,964 docs	Text → Text	Nutrition; General Medicine	Docs; Queries	Yes
TREC-COVID (Voorhees et al., 2021)	171,332 articles	Text → Text	COVID-19 Research	Articles; Queries	Yes
MIMIC-CXR (Johnson et al., 2019)	377,110 X-rays; 227,835 reports	Text ↔ Image	Chest Radiology	X-rays; Reports	Yes
ImageCLEFmed (Pelka et al., 2024)	∼66,000 images	Multi-query (Text/Image → Images)	Radiology; Pathology; Dermatology	X-ray, CT, MRI; Reports	Yes
3D-MIR (Abacha et al., 2023)	4 anatomies (Colon, Liver, Lung, Pancreas)	3D CT → Volume	Multi-organ Imaging	3D CT	Yes
BIMCV-R (Chen et al., 2024)	8,069 CT volumes	Text → Image	Respiratory (COVID; Pneumonia)	CT; Reports	Yes
CBIR (TotalSegmentator) (Li et al., 2021)	29 coarse; 104 detailed regions	Region-based (Segment → Scan)	Multi-organ Anatomy	Volumetric Scans	No
***M3Retrive* (Ours)**	∼ 164947 queries; ∼ 1238038 docs	Multimodal (Text+Image → Text/Image)	16 specialties (e.g., Anatomy, Cardio., Pulmo., Derm., Endo., Neuro., Radiol.)	Docs; Queries	Yes

Abbreviations: CT = computed tomography; MRI = magnetic resonance imaging.

work can be summarised as :

a) Introduction of *M3Retrieve*. We present the first large-scale multimodal retrieval benchmark for the medical domain. ***M3Retrieve*** accepts multimodal queries and targets realistic document stores spanning multiple specialties.

b) **Comprehensive Dataset.** ***M3Retrieve*** aggregates 22 manually-curated datasets (all under permissive licences) that cover **16 medical disciplines** and comprise **920 K text documents** plus **818 K multimodal queries**, providing broad coverage of real-world clinical scenarios.

c) Clinically-Grounded Task Suite. Guided by consultations with healthcare professionals, we define five retrieval tasks that mirror routine information-seeking workflows: **Visual Context Retrieval** (image + short text/caption → relevant passage), **Multimodal Query-to-Image Retrieval** (image or text description → visually similar image), **Case Study Retrieval** (image + patient transcript → closest full past case), and **Multimodal Summarisation Retrieval** (long report + associated images → concise summary).

d) Systematic Performance Evaluation We benchmark several state-of-the-art multimodal retrieval models on ***M3Retrieve***, revealing discipline-specific challenges and quantifying their impact on retrieval effectiveness.

2 Related Works

2.1 Retrieval Benchmarks

The evaluation of retrieval systems has a long and rich history. Early efforts, rooted in the Cranfield paradigm and later formalized by initiatives such as the Text Retrieval Conference (TREC) (Wikipedia contributors, 2024), established core evaluation measures such as precision, recall, and mean average precision—that continue to underpin retrieval performance assessment today. Over time, large-scale benchmarks like MS MARCO (Bajaj et al., 2018) and BEIR (Thakur et al., 2021) have provided standardized test collections and protocols for open-domain text retrieval, thereby driving the development of more robust retrieval models. In the medical domain, the unique nature of clinical language and the critical need for factual correctness have spurred the creation of specialized benchmarks. Initiatives such as BioASQ (Jeong et al., 2021), PubMedQA (Jin et al., 2019), and other medical question answering datasets (Ngo et al., 2024) have primarily focused on text retrieval and QA tasks, evaluating models on their ability to retrieve and reason over biomedical literature. However, these benchmarks rarely incorporate non-textual data even though medical diagnosis and decision support often require the interpretation of images (e.g., radiographs or histology slides) alongside text. Language-specific and domain-specific retrieval benchmarks have further refined evaluation criteria by addressing nuances in linguistic usage. For instance, initiatives like MIRACL (Zhang et al., 2022), mMARCO (Bonifacio et al., 2022), and various language-specific benchmarks (Acharya et al., 2024; Snegirev et al., 2025) have motivated the development of more effective multilingual retrieval models. Similarly, fine-grained domain-specific analyses in BEIR-like benchmarks have fostered the advancement of domain-agnostic embedding models. The most prominent and extensive multimodal retrieval benchmark to date is UniIR (Wei et al., 2023). However, the datasets within UniIR do not comprehensively address the intricate challenges of the medical domain. Medical applications demand fine-grained clinical detail, domain-specific terminologies, and the integration of both visual and textual evidence to support accurate decision-making and thus demand a separate benchmark of their own for a more holistic evaluation.

2.2 Retrieval Models

The evolution of retrieval systems in Natural Language Processing (NLP) has progressed from lexicon-based models to advanced dense and multimodal architectures (Ghosh et al., 2024c). Early retrieval models like BM-25 (Robertson and Zaragoza, 2009) relied on lexicons. With the rise of deep learning, dense retrieval models, such as E5 (Wang et al., 2022), BGE (Xiao et al., 2023), and NV Embed (Lee et al., 2025), have addressed vocabulary mismatch by utilizing contrastive learning to produce semantic embeddings. The integration of multiple modalities began with CLIP (Radford et al., 2021), aligning visual and textual representations in a shared embedding space for cross-modal retrieval. This approach was extended in models like MM Ret (Zhou et al., 2024) and MedImageInsight (Codella et al., 2024), which specialize in medical image-text retrieval. Unified multimodal retrievers such as the UniIR family (Wei et al., 2023)(CLIP SF and BLIP FF) enable cross-modal retrieval across diverse data types. Recent models like VLM2Vec (Jiang et al., 2025) and MM Embed (Lin et al., 2024) further improve joint representation learning for text and images. Additionally, late-interaction models like FLMR (Lin et al., 2023) compute token-level similarities for more precise retrieval relevance determination. We believe that with the introduction of *M3Retrieve*, the community will be better equipped to assess discipline-specific challenges and the integration of multimodal signals, thereby helping in establishing practical guidelines for developing more reliable medical retrieval systems.

Task	#Corpus	# Queries
Visual Context Retrieval	507101	93488
Summary Retrieval	228887	3015
Query to Image Retrieval	2050	671
Case study Retrieval	500000	67773
Total	1238038	164947

Table 2: Statistics of the Dataset in the *M3Retrieve* Benchmark showing the number of corpus and query in the evaluation set for each task in the benchmark.

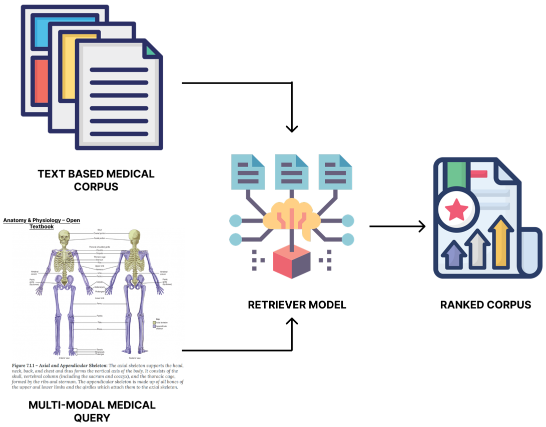

TEXT BASED MEDICAL CORPUS
Anatomy & Physiology – Open Textbook
Figure 7.1.1 – Axial and Appendicular Skeleton: The axial skeleton supports the head, neck, back, and chest and thus forms the vertical axis of the body. It consists of the skull, vertebral column (including the sacrum and coccyx), and the thoracic cage, formed by the ribs and sternum. The appendicular skeleton is made up of all bones of the upper and lower limbs and the girdles which attach them to the axial skeleton.
MULTI-MODAL MEDICAL QUERY
RETRIEVER MODEL
RANKED CORPUS

Figure 2: Overview of a retrieval task addressed in the *M3Retrieve* Benchmark. The task aims to integrate both text and image data, with the retriever model ranking documents based on relevance. The multimodal framework enriches retrieval performance by incorporating visual information alongside traditional text-based retrieval.

We present ***M3Retrieve***, a comprehensive multimodal medical retrieval benchmark for healthcare domain that consists of 5 different tasks. In this section, we provide a detailed formulation for each of the tasks decided, outline the diverse data sources utilized in constructing the ***M3Retrieve***, and describe the methodology employed to curate the benchmark from these sources.

3 Creation of the *M3Retrieve* Benchmark

The following section outlines the tasks included in the *M3Retrieve* Benchmark and describes the methodology used for their construction. Table 2 gives an overview of the number of corpus and queries for each of the tasks in the *M3Retrieve* Benchmark.

3.1 VisualContext Retrieval

Task Formulation: Given a multimodal query $Q = (Q_{\text{text}}, Q_{\text{image}})$ and a retrieval corpus $D = \{D_1, D_2, \ldots, D_n\}$, the retriever aims to find and rank the set of relevant documents:

$$D_Q^+ = \{D_{Q,1}^+, \ldots, D_{Q,m}^+\} \subset D, \quad \text{where } m \ll n$$

where, **Positive documents** D_Q^+ represents directly relevant documents containing information aligned with the query's text and image.

3.1.1 Data Curation and Relevance Design

Data Sources: For the VisualContext Retrieval task we collect data from three primary open-access sources namely: **(i)** Wikipedia, using the Wikimedia API[5] to extract structured multimodal content across diverse medical domains, yielding 288,983 corpus documents and 24,523 multimodal queries; **(ii)** PubMed, leveraging a subset of 10,000 articles from the dataset by Li et al. (Li et al., 2023), resulting in 204,217 corpus documents and 64,366 text-image queries; and **(iii)** Open Access medical textbooks sourced from Creative Commons licensed platforms such as OpenStax and Open Oregon State, processed using similar extraction and pairing strategies.

Relevancy Mapping Formulation: Each data source presented unique challenges, so we collaborated with medical experts to define dataset-specific relevancy mappings. For Wikipedia, experts noted that images are generally relevant to the entire article, making it difficult to distinguish between related and unrelated paragraphs. Consequently, for Wikipedia, we defined the relevance mapping as follows:

Relevancy Mapping (Wikipedia)

Strategy: Images explicitly referred to in a paragraph were assigned a relevance score of **2**. All other paragraphs within the same article received a score of **1** due to their contextual relation to the overall topic.

We used 10,000 PubMed articles from (Li et al., 2023) and, based on expert input, assigned the highest relevance to the paragraph referencing a figure, as its significance diminishes elsewhere

Relevancy Mapping (PubMed)

Strategy: The paragraph that explicitly mentions the figure was assigned a score of **2**, while all other paragraphs were excluded from consideration. This reflects the highly localized relevance of figures in scientific articles.

In textbooks, each image-caption pair forms a multimodal query, with paragraphs as candidate documents; figures are typically relevant only within their specific textbook section.

Relevancy Mapping(TextBooks)

Strategy: The paragraph that directly references a figure was assigned a score of **2**. All other paragraphs within the same section were considered contextually relevant and given a score of **1**.

This unified scoring framework across data sources allowed the creation of a consistent, high-quality benchmark for evaluating multimodal evidence retrieval systems.

3.2 Multimodal Summary Retrieval

Task Formulation. Given a multimodal context $Q = (Q_{\text{text}},\ Q_{\text{image}})$, where Q_{text} may represent a clinical note, patient conversation, or textual report, and Q_{image} is an associated medical image (e.g., X-ray, MRI), the goal is to retrieve the most relevant summary from a candidate summary pool

$$S = \{S_1, S_2, \ldots, S_n\},$$

where each S_i is a standalone summary. Exactly one candidate $S_Q^\star$ is considered the correct summary, which best captures the key information from both modalities.

3.2.1 Data Curation and Relevance Design

Data Sources. We use the MMQS dataset (Ghosh et al., 2024a) for the task of multimodal healthcare summarization. MMQS consists of 3,015 curated samples, where each instance comprises a multimodal medical query—combining a textual component (e.g., patient query or clinical dialogue) and a visual component (e.g., a relevant medical image)—paired with a corresponding summary that integrates information from both modalities. The summaries are drawn from a larger retrieval corpus derived from HealthcareMagic (Mrini et al.,

[5] https://api.wikimedia.org/wiki/Getting_started_with_Wikimedia_APIs

2021), which contains 228,887 medical queries and their corresponding expert-written summaries. This corpus serves as the foundation for candidate summaries during retrieval, providing a rich pool of medical knowledge spanning both textual and visual contexts. Here, we augment the original HealthcareMagic dataset (Mrini et al., 2021) by appending the curated multimodal summaries from MMQS, thereby constructing an expanded summarization corpus for retrieval.

Relevancy Mapping Formulation

Each multimodal medical query serves as the input, and the retriever is tasked with identifying the corresponding summary from a large-scale summarization corpus derived from HealthcareMagic (Mrini et al., 2021).

Relevancy Mapping

Strategy: The summary that directly corresponds to the given multimodal query is assigned a score of 2, while all other summaries in the corpus are assigned a score of 0.

3.3 Multimodal Query to Image Retrival

Task Formulation. Given a multimodal query $Q = (Q_{\text{text}}, Q_{\text{image}})$, where Q_{text} represents a textual medical query or dialogue and Q_{image} provides visual context (e.g., an indicative or reference image), the task is to retrieve the most relevant image from a candidate image pool

$$I = \{I_1, I_2, \ldots, I_n\},$$

where each I_i is a standalone medical image (e.g., chest X-ray, MRI, ultrasound, pathology scan). Exactly one candidate image $I_Q^\star$ is considered the correct or best matching image for the query.

3.3.1 Data Curation and Relevance

Data Sources: To support the novel retrieval task of selecting relevant medical images based on a multimodal input query (consisting of textual and visual cues), we curated a dataset derived from the public MedPix 2.0 (Siragusa et al., 2024) repository—a comprehensive radiology teaching file provided by the U.S. National Library of Medicine. MedPix 2.0 is the best dataset for this task because it offers expertly curated, semantically rich text–image pairs across diverse medical conditions, enabling precise and clinically grounded multimodal query-to-image relevance mapping.

Relevancy Mapping Formulation: To establish reliable ground-truth for the multimodal image query → image retrieval task, we leverage the structural integrity of the MedPix 2.0 dataset, where each clinical case is identified by a unique `U_id`. All images associated with the same `U_id` are considered *relevant* to the query composed from that case's textual description and accompanying visual clue.

Relevancy Mapping

Strategy: All images that share the same `U_id` as the multimodal query (i.e., derived from the same clinical case) are assigned a score of 2 and all other images in the corpus are assigned a score of 0.

3.4 Case Study Retrieval

Task Formulation. Given a multimodal clinical query $Q = (Q_{\text{text}}, Q_{\text{image}})$, where Q_{text} represents a patient complaint, diagnostic note, or medical dialogue, and Q_{image} is an associated clinical image (e.g., scan, X-ray, pathology slide), the goal is to retrieve the most relevant case study from a set of documented medical cases

$$S = \{S_1, S_2, \ldots, S_n\},$$

where each S_i is a structured medical case study consisting of textual findings, diagnoses, and possibly images. Exactly one case study $S_Q^\star$ is most relevant to the query.

3.4.1 Data Curation and Relevance

Data Sources. To enable the retrieval of clinically relevant case studies based on multimodal (Image + Textual Description), we make use of the MultiCaRe (Nievoff, 2024) dataset—a publicly available resource constructed from open-access case reports on PubMed Central. This dataset offers a comprehensive collection of over 93,000 de-identified clinical cases paired with more than 130,000 diagnostic images. Its broad medical coverage and well-aligned text–image pairs make it an ideal foundation for developing and evaluating multimodal retrieval systems in real-world healthcare settings.

Relevance Mapping Formulation To establish reliable ground-truth for the multimodal relevance mapping task, we leverage the structured design of the MultiCaRe dataset, where each clinical case is tagged by a unique case_id. All textual narratives and associated images sharing the same case_id are considered relevant to any query derived from that case's combined textual and visual information.

Relevancy Mapping

Strategy: All case studies that share the same `case_id` as the multimodal query (i.e., originate from the same clinical record) are assigned a score of 2. Case studies with similar diagnostic categories or overlapping symptoms but from different records are assigned a score of 1, while all remaining case studies in the corpus are assigned a score of 0.

3.5 Quality Control Using Domain Expert

Throughout the data curation process, medical experts provided valuable feedback to ensure the selection of the most relevant data sources and essential medical modalities, enhancing the dataset's quality and applicability. Their insights were instrumental in establishing accurate relevance mappings, ensuring that query-document relationships aligned with real-world medical reasoning. Additionally, to validate the dataset's reliability, a sample of 80 queries across each task was reviewed by two doctors. The evaluation yielded a *Cohen's kappa score of 0.78*, indicating a high level of agreement between the reviewers and confirming that the dataset rankings were accurate and meaningful.

4 Experimental Setup

All experiments were conducted using the MTEB Python library [6] on NVIDIA A100 80GB GPUs. FLMR was evaluated using the implementation available at `https://github.com/LinWeizheDragon/FLMR`, and document retrieval was performed using BM-25 via Pyserini [7]. For both FLMR and BM-25, the evaluation metrics were computed using the pytrec_eval [8] Python library, following the implementation in the MTEB library. We used nNDCG@10 as the primary metric for evaluation.

4.1 Baseline Models

To evaluate retrieval performance on the ***M3Retrieve***Benchmark, we assess a range of uni-modal and multimodal models, categorized as follows:

- **Lexicon-Based Model: BM25** (Robertson and Zaragoza, 2009) — A strong traditional baseline using term frequency and inverse document frequency for scoring.

- **Text-Based Encoders: E5-Large-v2** (Wang et al., 2022) (1024-dim; weakly-supervised contrastive learning), **BGE-en-Large** (Xiao et al., 2023) (1024-dim; top MTEB performance), **NV-Embed-v2** (Lee et al., 2025) (4096-dim; MTEB leader with latent-attention pooling).

- **CLIP-Style Models: MMRet-Large** (Zhou et al., 2024) (CLIP-based; 768-dim; context length 77), **MedImageInsight (MII)** (Codella et al., 2024) (medical domain; CLIP-style contrastive learning), **CLIP-SF** (Wei et al., 2023) (768-dim; context length 77).

- **Multimodal Encoders: BLIP-FF** (Wei et al., 2023) (BLIP-based; 768-dim; context length 512), **MM-Embed** (Lin et al., 2024) (extends NV-Embed-v1; state-of-the-art on UniIR and MTEB).

- **Multimodal Late Interaction Retriever: FLMR** (Lin et al., 2023) — Uses token-level similarity for fine-grained late interaction between queries and documents.

5 Results Analysis

To assess the effectiveness of various retrieval models in the medical domain, we evaluated multiple uni-modal and multimodal approaches on the ***M3Retrieve***. The models analyzed can be divided into five broad categories: **lexicon-based**, **text-based dense encoders**, **CLIP-style multimodal retrievers**, **multimodal encoders**, and **late-interaction models**.

Table 3 presents the NDCG@10 scores for various retrieval models across the four tasks in the ***M3Retrieve***Benchmark. The models are categorized into five retrieval paradigms: lexicon-based retrievers, uni-modal dense retrievers, CLIP-style models, multi-modal dense retrievers, and late-interaction multi-modal retrievers.

It should also be noted that the reported results are grounded in the structural assumptions of the underlying datasets.

5.1 Overall Performance Trends

Multimodal models exhibit significant potential, particularly in tasks that inherently require the in-

[6] `https://github.com/embeddings-benchmark/mteb`

[7] `https://github.com/castorini/pyserini`

[8] `https://github.com/cvangysel/pytrec_eval`

Method	VisualContext Retrieval	Summary Retrieval	Query to Image Retrieval	Case Study Retrieval
BM-25	38.07	18.16	N/A	**11.50**
E5 Large	35.14	70.23	N/A	7.68
BGE	32.32	83.66	N/A	6.59
NV Embed	43.28	**89.73**	N/A	10.99
MM Ret	24.56	43.71	2.27	1.09
MII	28.13	22.5	**43.53**	1.64
CLIP SF	26.44	26.30	29.06	1.27
BLIP FF	24.72	20.89	2.23	0.92
MM Embed	**45.47**	76.27	29.49	9.91
FLMR	24.80	21.30	2.56	1.48

Table 3: NDCG@10 scores for ten retrieval models representing different retrieval styles, including lexicon-based retrievers, uni-modal dense retrievers, CLIP-style models, multi-modal dense retrievers, and late-interaction multi-modal retrievers on the ***M3Retrieve***Benchmark. The best-performing model has been highlighted as **bold** while the second-best model has been underlined.

tegration of textual and visual information. For instance, in the **VisualContext Retrieval** task, the **MM-Embed** model achieves the highest NDCG@10 score of 45.47, outperforming all other models. Similarly, in the **Query to Image Retrieval** task, the **MedImageInsight** model leads with a score of 43.53. These results underscore the advantage of multimodal models in scenarios where both text and image modalities are crucial.

However, in tasks that are predominantly textual, such as **Summary Retrieval** and **Case Study Retrieval**, uni-modal dense retrievers demonstrate superior performance. The **NV-Embed** model achieves the highest scores in both tasks, with 89.73 and 10.99, respectively. This suggests that, in the current landscape, uni-modal models remain highly effective for text-centric retrieval tasks.

5.2 Task-Specific Model Performance

VisualContext Retrieval: This task benefits from models capable of integrating multimodal information. The **MM-Embed** model achieves the highest performance (45.47), followed by the **NV-Embed** (43.28) and **BM25** (38.07). The strong performance of **MM-Embed** highlights the effectiveness of multimodal dense retrievers in capturing the nuanced relationships between text and images.

Summary Retrieval: Uni-modal dense retrievers dominate this task, with **NV-Embed** achieving the top score of 89.73, followed by **BGE** (83.66) and **E5 Large** (70.23). Multimodal models lag behind, indicating that current multimodal approaches may not yet effectively handle tasks that are primarily textual.

Query to Image Retrieval: The **MedImageInsight** model leads with a score of 43.53, demonstrating the strength of CLIP-style models in image retrieval tasks. The **MM-Embed** model follows with 29.49, and **CLIP SF** achieves 29.06. These results suggest that models trained with contrastive learning on image-text pairs are particularly effective for image-centric retrieval tasks.

Case Study Retrieval: The **NV-Embed** model again achieves the highest score (10.99), indicating that uni-modal dense retrievers are currently more effective for retrieving comprehensive case studies. The **MM Ret** model follows closely with 10.87, suggesting some potential for multimodal models in this area.

6 Conclusion

We introduce ***M3Retrieve***, the first comprehensive multimodal Medical Retrieval Benchmark, designed to evaluate retrieval models across 16+ medical domains, covering 500K+ documents and 100K queries. This benchmark rigorously assesses uni-modal (text-based) and multimodal retrieval models across five retrieval styles and architectural approaches, providing a detailed analysis of their performance in complex medical scenarios. Given the critical role of accurate medical information retrieval in patient care, clinical decision-making, and research, this benchmark helps identify gaps and strengths in current models. By benchmarking a diverse range of retrieval approaches, ***M3Retrieve***establishes a strong foundation for developing more effective and specialized multimodal retrieval systems. We believe it will be instrumental in driving future research and advancing medical AI systems to enhance real-world healthcare applications.

7 Ethical Considerations

All data in ***M3Retrieve***comes from publicly available sources with Creative Commons (CC) licenses, ensuring compliance with HIPAA and GDPR. Medical professionals were involved throughout the dataset design to ensure relevance and accuracy. A human evaluation was conducted post-construction to verify quality and fairness, reinforcing ethical and responsible medical benchmark.

8 Limitation and Future Work

The ***M3Retrieve***is a foundational step toward a diverse multi-domain, multimodal medical retrieval benchmark, but it has certain limitations. Currently, it covers 16 broad medical domains, which may not encompass all specialties. Future work will expand coverage to additional disciplines. The benchmark currently features multimodal queries with text-only documents, a common setup, but future versions will incorporate more modalities for both queries and corpora. Our findings show that existing models underperform in medical retrieval compared to general-domain tasks, underscoring the need for a medical-specific multimodal retrieval model. Additionally, our figure-to-paragraph mappings were validated only through a limited manual review, and large-scale verification across the full dataset was not feasible, making our approach reliant on the assumption that figures are responsibly referenced in the original publications.

9 Acknowledgement

We thank the ACL ARR reviewers for their constructive comments and valuable feedback on the draft. Akash Ghosh and Sriparna Saha extend their sincere appreciation to the SERB (Science and Engineering Research Board) POWER scheme, Department of Science and Engineering, Government of India of India, for generously funding this research.

References

Asma Ben Abacha, Alberto Santamaria-Pang, Ho Hin Lee, Jameson Merkow, Qin Cai, Surya Teja Devarakonda, Abdullah Islam, Julia Gong, Matthew P. Lungren, Thomas Lin, Noel C Codella, and Ivan Tarapov. 2023. 3d-mir: A benchmark and empirical study on 3d medical image retrieval in radiology. *Preprint*, arXiv:2311.13752.

Arkadeep Acharya, Rudra Murthy, Vishwajeet Kumar, and Jaydeep Sen. 2024. Benchmarking and building zero-shot hindi retrieval model with hindi-beir and nllb-e5. *Preprint*, arXiv:2409.05401.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. Ms marco: A human generated machine reading comprehension dataset. *Preprint*, arXiv:1611.09268.

Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2022. mmarco: A multilingual version of the ms marco passage ranking dataset. *Preprint*, arXiv:2108.13897.

Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. 2016. A full-text learning to rank dataset for medical information retrieval. In *Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20–23, 2016. Proceedings 38*, pages 716–722. Springer.

Yinda Chen, Che Liu, Xiaoyu Liu, Rossella Arcucci, and Zhiwei Xiong. 2024. Bimcv-r: A landmark dataset for 3d ct text-image retrieval. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 124–134. Springer.

Noel C. F. Codella, Ying Jin, Shrey Jain, Yu Gu, Ho Hin Lee, Asma Ben Abacha, Alberto Santamaria-Pang, Will Guyman, Naiteek Sangani, Sheng Zhang, Hoifung Poon, Stephanie Hyland, Shruthi Bannur, Javier Alvarez-Valle, Xue Li, John Garrett, Alan McMillan, Gaurav Rajguru, Madhu Maddi, and 12 others. 2024. Medimageinsight: An open-source embedding model for general domain medical imaging. *Preprint*, arXiv:2410.06542.

Akash Ghosh, Arkadeep Acharya, Raghav Jain, Sriparna Saha, Aman Chadha, and Setu Sinha. 2024a. Clipsyntel: clip and llm synergy for multimodal question summarization in healthcare. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22031–22039.

Akash Ghosh, Arkadeep Acharya, Prince Jha, Sriparna Saha, Aniket Gaudgaul, Rajdeep Majumdar, Aman Chadha, Raghav Jain, Setu Sinha, and Shivani Agarwal. 2024b. Medsumm: A multimodal approach to summarizing code-mixed hindi-english clinical queries. In *European Conference on Information Retrieval*, pages 106–120. Springer.

Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024c. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. *arXiv preprint arXiv:2404.07214*.

Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Gaurav Pandey, Dinesh Raghu, and Setu Sinha. 2024d. Healthalignsumm: Utilizing alignment for multimodal summarization of code-mixed healthcare dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11546–11560.

Minbyul Jeong, Mujeen Sung, Gangwoo Kim, Donghyeon Kim, Wonjin Yoon, Jaehyo Yoo, and Jaewoo Kang. 2021. Transferability of natural language inference to biomedical question answering. *Preprint*, arXiv:2007.00217.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. 2025. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. *Preprint*, arXiv:2410.05160.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025. Nv-embed: Improved techniques for training llms as generalist embedding models. *Preprint*, arXiv:2405.17428.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Preprint*, arXiv:2306.00890.

Xiaoqing Li, Jiansheng Yang, and Jinwen Ma. 2021. Recent developments of content-based image retrieval (cbir). *Neurocomputing*, 452:675–689.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. 2024. Mm-embed: Universal multimodal retrieval with multimodal llms. *Preprint*, arXiv:2411.02571.

Weizhe Lin, Jinghong Chen, Jingbiao Mei, Alexandru Coca, and Bill Byrne. 2023. Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. *Preprint*, arXiv:2309.17133.

Haohao Luo, Ying Shen, and Yang Deng. 2023. Unifying text, tables, and images for multimodal question answering. Association for Computational Linguistics.

Khalil Mrini, Franck Dernoncourt, Walter Chang, Emilia Farcas, and Ndapandula Nakashole. 2021. Joint summarization-entailment optimization for consumer health question understanding. In *Proceedings of the second workshop on natural language processing for medical conversations*, pages 58–65.

Nghia Trung Ngo, Chien Van Nguyen, Franck Dernoncourt, and Thien Huu Nguyen. 2024. Comprehensive and practical evaluation of retrieval-augmented generation systems for medical question answering. *Preprint*, arXiv:2411.09213.

Mauro Nievoff. 2024. Multicare dataset. `https://github.com/mauro-nievoff/MultiCaRe_Dataset`. Accessed: 2025-05-17.

Obioma Pelka, Asma Ben Abacha, Alba García Seco de Herrera, Jitpakorn Jacutprakart, Henning Müller, and Yashin Dicente Cid. 2024. Overview of ImageCLEFmedical 2024 – caption prediction and concept detection. In *CLEF 2024 Working Notes*, CEUR Workshop Proceedings, Grenoble, France. CEUR-WS.org.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Irene Siragusa, Salvatore Contino, Massimo La Ciura, Rosario Alicata, and Roberto Pirrone. 2024. Medpix 2.0: a comprehensive multimodal biomedical dataset for advanced ai applications. *arXiv preprint arXiv:2407.02994*.

Artem Snegirev, Maria Tikhonova, Anna Maksimova, Alena Fenogenova, and Alexander Abramov. 2025. The russian-focused embedders' exploration: rumteb benchmark and russian embedding model design. *Preprint*, arXiv:2408.12503.

Lin Sun, Kai Zhang, Qingyuan Li, and Renze Lou. 2024. Umie: Unified multimodal information extraction with instruction tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19062–19070.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir:

A heterogenous benchmark for zero-shot evaluation of information retrieval models. *Preprint*, arXiv:2104.08663.

Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2021. Trec-covid: constructing a pandemic information retrieval test collection. In *ACM SIGIR Forum*, volume 54, pages 1–12. ACM New York, NY, USA.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Tianshi Wang, Fengling Li, Lei Zhu, Jingjing Li, Zheng Zhang, and Heng Tao Shen. 2025. Cross-modal retrieval: a systematic review of methods and future directions. *Proceedings of the IEEE*.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. 2023. Uniir: Training and benchmarking universal multimodal information retrievers. *Preprint*, arXiv:2311.17136.

Wikipedia contributors. 2024. Text retrieval conference — Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Text_Retrieval_Conference. Accessed: 2024-02-11.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2022. Making a miracl: Multilingual information retrieval across a continuum of languages. *Preprint*, arXiv:2210.09984.

Junjie Zhou, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, Defu Lian, and Yongping Xiong. 2024. Megapairs: Massive data synthesis for universal multimodal retrieval. *Preprint*, arXiv:2412.14475.

A Appendix

A.1 Example of datapoints in the *M3Retrieve Benchmark*

Figure 3 exhibits an example from the ***M3Retrieve*** Benchmark showing the query image-text pair and corpus texts along with justifications for the assigned scores.

A.2 Discipline wise analysis of the tasks in the *M3Retrieve* Benchmark

A.2.1 Visual Context Retrieval

Based on the results in Table 4 we can make the following conclusions :

1) Anatomy and Physiology: NV Embed (66.29) and BM25 (61.20) outperform all other models, showing that **dense text retrieval remains the most effective method for this domain**.

2) Psychiatry and Mental Health: MM Embed (50.28) is the strongest performer, demonstrating the benefit of **multimodal representation learning** in capturing mental health-related contexts.

3) PubMed Retrieval: MM Embed (70.06) surpasses **BM25 (68.72)**, highlighting that multimodal encoders provide a **more comprehensive representation** in large-scale biomedical literature retrieval.

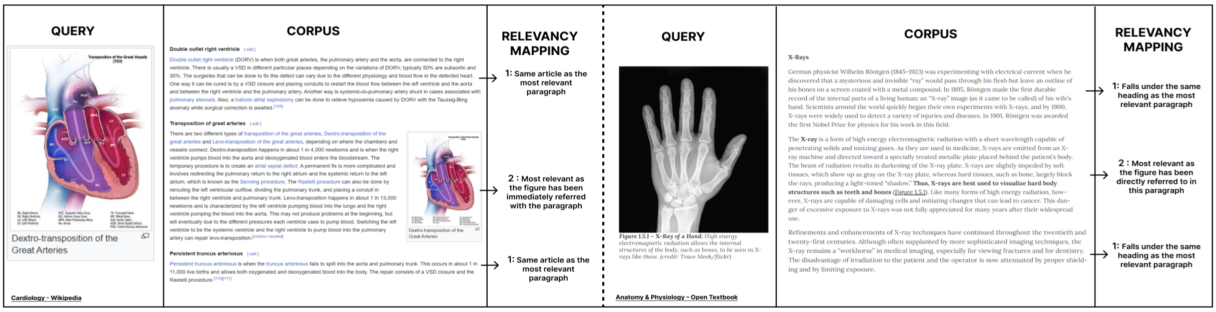
4) Orthopedics and Musculoskeletal: Scores remain **low across all models** (BM25: 6.27, NV Embed: 6.86), indicating **retrieval challenges** in this sub-domain, possibly due to **complex terminologies and limited training data**.

Overall, **dense encoders** (MM Embed, NV Embed) outperform both **BM25** and **CLIP-style models**, highlighting the benefits of **deep learning-based joint embeddings**. **BM25 remains a strong baseline**, particularly in text-heavy disciplines.

The superior performance of MM Embed over all unimodal models, including NV Embed, the best-performing text retrieval model, emphasizes the **importance of multimodal representation learning** for medical retrieval in *M3Retrieve*. However, NV Embed and MM Embed show **30% and 25.37% lower** NDCG@10 scores than their BEIR averages (62.65 and 60.3).

A.2.2 Case Study Retrieval

According to Table 5, across the fifteen medical specialties, traditional term-matching via BM25 remains a strong baseline, especially in Orthopedics (14.21), Pathology (13.51), and Gastroenterology (13.08). Neural text encoders (E5 and BGE) demonstrate particular strength in Gastroenterology (E5 = 8.84, BGE = 7.99) and Genetics and Genomics (E5 = 8.37, BGE = 7.13), suggesting their aptitude for capturing nuanced biomedical language. The NV Embed model shows a clear niche in Genetics and Genomics (13.59), while multimodal retrieval (MM Ret) and the Medical


QUERY
Transposition of the Great Vessels (TGV)
SVC
Ao
PDA
MPA
LA
RA
ASD
AoV
PV
MV
TV
LV
RV
IVC
RA. Right Atrium
RV. Right Ventricle
LA. Left Atrium
LV. Left Ventricle
SVC. Superior Vena Cava
IVC. Inferior Vena Cava
MPA. Main Pulmonary Artery
Ao. Aorta
TV. Tricuspid Valve
MV. Mitral Valve
AoV. Aortic Valve
ASD. Atrial Septal Defect
PDA. Patent Ductus Arteriosis
Dextro-transposition of the Great Arteries
Cardiology - Wikipedia
CORPUS
RELEVANCY MAPPING
1: Same article as the most relevant paragraph
2 : Most relevant as the figure has been immediately referred with the paragraph
1: Same article as the most relevant paragraph
QUERY
Anatomy & Physiology – Open Textbook
CORPUS
RELEVANCY MAPPING
1: Falls under the same heading as the most relevant paragraph
2 : Most relevant as the figure has been directly referred to in this paragraph
1: Falls under the same heading as the most relevant paragraph

Double outlet right ventricle [edit]

Double outlet right ventricle (DORV) is when both great arteries, the pulmonary artery and the aorta, are connected to the right ventricle. There is usually a VSD in different particular places depending on the variations of DORV, typically 50% are subaortic and 30%. The surgeries that can be done to fix this defect can vary due to the different physiology and blood flow in the defected heart. One way it can be cured is by a VSD closure and placing conduits to restart the blood flow between the left ventricle and the aorta and between the right ventricle and the pulmonary artery. Another way is systemic-to-pulmonary artery shunt in cases associated with pulmonary stenosis. Also, a balloon atrial septostomy can be done to relieve hypoxemia caused by DORV with the Taussig-Bing anomaly while surgical correction is awaited.[109]

Transposition of great arteries [edit]

There are two different types of transposition of the great arteries, Dextro-transposition of the great arteries and Levo-transposition of the great arteries, depending on where the chambers and vessels connect. Dextro-transposition happens in about 1 in 4,000 newborns and is when the right ventricle pumps blood into the aorta and deoxygenated blood enters the bloodstream. The temporary procedure is to create an atrial septal defect. A permanent fix is more complicated and involves redirecting the pulmonary return to the right atrium and the systemic return to the left atrium, which is known as the Senning procedure. The Rastelli procedure can also be done by rerouting the left ventricular outflow, dividing the pulmonary trunk, and placing a conduit in between the right ventricle and pulmonary trunk. Levo-transposition happens in about 1 in 13,000 newborns and is characterized by the left ventricle pumping blood into the lungs and the right ventricle pumping the blood into the aorta. This may not produce problems at the beginning, but will eventually due to the different pressures each ventricle uses to pump blood. Switching the left ventricle to be the systemic ventricle and the right ventricle to pump blood into the pulmonary artery can repair levo-transposition.[*citation needed*]

Dextro-transposition of the Great Arteries

Persistent truncus arteriosus [edit]

Persistent truncus arteriosus is when the truncus arteriosus fails to split into the aorta and pulmonary trunk. This occurs in about 1 in 11,000 live births and allows both oxygenated and deoxygenated blood into the body. The repair consists of a VSD closure and the Rastelli procedure.[110][111]

Figure 1.5.1 – X-Ray of a Hand: *High energy electromagnetic radiation allows the internal structures of the body, such as bones, to be seen in X-rays like these. (credit: Trace Meek/flickr)*

X-Rays

German physicist Wilhelm Röntgen (1845–1923) was experimenting with electrical current when he discovered that a mysterious and invisible "ray" would pass through his flesh but leave an outline of his bones on a screen coated with a metal compound. In 1895, Röntgen made the first durable record of the internal parts of a living human: an "X-ray" image (as it came to be called) of his wife's hand. Scientists around the world quickly began their own experiments with X-rays, and by 1900, X-rays were widely used to detect a variety of injuries and diseases. In 1901, Röntgen was awarded the first Nobel Prize for physics for his work in this field.

The **X-ray** is a form of high energy electromagnetic radiation with a short wavelength capable of penetrating solids and ionizing gases. As they are used in medicine, X-rays are emitted from an X-ray machine and directed toward a specially treated metallic plate placed behind the patient's body. The beam of radiation results in darkening of the X-ray plate. X-rays are slightly impeded by soft tissues, which show up as gray on the X-ray plate, whereas hard tissues, such as bone, largely block the rays, producing a light-toned "shadow." **Thus, X-rays are best used to visualize hard body structures such as teeth and bones** (Figure 1.5.1). Like many forms of high energy radiation, however, X-rays are capable of damaging cells and initiating changes that can lead to cancer. This danger of excessive exposure to X-rays was not fully appreciated for many years after their widespread use.

Refinements and enhancements of X-ray techniques have continued throughout the twentieth and twenty-first centuries. Although often supplanted by more sophisticated imaging techniques, the X-ray remains a "workhorse" in medical imaging, especially for viewing fractures and for dentistry. The disadvantage of irradiation to the patient and the operator is now attenuated by proper shielding and by limiting exposure.

Figure 3: An example from the *M3Retrieve* Benchmark showing the query image-text pair and corpus texts along with justifications for the assigned scores.

Discipline	BM-25	E5	BGE	NV Embed	MM Ret	MII	CLIP SF	BLIP FF	MM Embed	FLMR
Anatomy and Physiology	61.20	49.17	60.15	**66.29**	47.25	49.15	47.96	52.91	<u>64.14</u>	2.05
Cardiology	39.25	33.26	31.15	<u>42.16</u>	26.12	27.56	27.66	26.32	**45.3**	32.02
Dermatology	41.78	42.26	34.10	<u>51.53</u>	33.37	36.27	34.70	33.49	**52.59**	35.05
Endocrinology and Diabetes	35.80	36.99	33.86	<u>45.83</u>	28.95	29.83	27.97	30.45	**46.38**	32.86
Gastroenterology	35.16	36.90	31.67	**46.07**	26.47	29.41	27.96	27.68	<u>46.01</u>	33.20
Hematology	32.23	31.89	29.00	<u>38.46</u>	22.65	24.01	23.65	20.53	**39.73**	27.47
Microbiology and Cell Biology	34.01	30.65	29.23	<u>39.92</u>	19.14	19.35	18.64	19.14	**39.80**	6.55
Miscellaneous	34.25	31.29	29.13	<u>39.76</u>	21.05	21.12	21.07	20.89	**41.93**	25.69
Neurology and Neuroscience	24.54	23.15	19.88	<u>28.04</u>	15.94	17.38	17.03	15.41	**30.33**	17.82
Ophthalmology and Sensory Systems	33.29	31.45	25.10	<u>38.98</u>	22.46	24.66	23.78	22.20	**41.86**	25.18
Orthopedics and Musculoskeletal	6.27	5.99	5.25	<u>6.86</u>	4.77	5.62	4.57	5.57	**12.31**	5.38
Pharmacology	37.47	35.19	29.42	<u>41.86</u>	25.59	27.52	26.99	25.01	**44.43**	30.80
Psychiatry and Mental Health	39.71	39.69	28.71	<u>40.74</u>	34.21	28.06	33.99	29.43	**50.28**	35.40
Pubmed	<u>68.72</u>	56.90	58.72	67.32	29.90	37.75	29.47	22.19	**70.06**	41.53
Radiology and Imaging	47.39	40.43	41.84	<u>51.11</u>	30.95	36.64	19.80	26.58	**52.86**	32.25
Reproductive System	32.05	32.25	25.87	<u>39.40</u>	25.09	28.28	29.29	24.48	**42.09**	28.66
Respiratory and Pulmonology	44.15	39.25	34.76	<u>49.31</u>	28.89	31.80	31.56	26.25	**50.43**	2.74
Surgical Specialties	38.07	36.10	34.06	<u>46.42</u>	25.42	32.23	27.90	26.93	**48.08**	31.71
Average	38.07	35.00	32.00	<u>43.00</u>	26.00	28.00	26.00	25.00	**45.00**	24.80

Table 4: NDCG@10 scores for ten retrieval models representing different retrieval styles, including lexicon-based retrievers, uni-modal dense retrievers, CLIP-style models, multi-modal dense retrievers, and late-interaction multi-modal retrievers for the Visual Context Retrieval task in the *M3Retrieve* Benchmark. The best-performing model has been highlighted as **bold** while the second best model has been <u>underlined</u>.

Discipline	BM-25	E5	BGE	NV Embed	MM Ret	MII	CLIP SF	BLIP FF	MM Embed	FLMR
Obstetrics and Gynecology	13.90	11.14	9.34	8.12	1.69	1.65	1.10	0.80	14.47	1.55
Hematology	10.32	6.33	5.62	9.02	1.16	1.06	1.04	1.05	7.84	0.80
Cardiology	10.09	6.76	6.92	9.88	0.98	2.31	2.08	0.58	8.98	2.30
Neurology	11.16	7.82	6.11	9.55	0.98	2.02	1.35	0.90	8.84	1.35
Orthopedics	14.21	10.39	9.09	11.23	1.29	2.70	1.50	1.10	9.90	0.95
Pathology	13.51	6.96	6.17	12.67	0.87	1.87	1.20	0.85	10.05	2.10
Endocrinology	11.14	7.53	5.73	9.97	0.95	1.39	0.81	1.01	10.01	1.40
Oncology	10.16	6.10	5.33	13.04	0.75	1.43	0.90	0.70	7.34	0.60
Pulmonology	11.07	6.88	6.54	8.98	0.90	1.43	1.05	1.00	9.03	1.45
Psychiatry and Behavioral Health	7.18	5.58	4.18	11.98	0.51	0.86	1.40	0.95	8.48	1.10
Genetics and Genomics	12.33	8.37	7.13	13.59	1.58	1.66	1.13	0.88	12.13	2.45
Infectious Diseases	12.08	8.06	6.45	11.48	1.22	1.31	1.18	0.92	9.90	0.75
Gastroenterology	13.08	8.84	7.99	12.11	1.30	1.95	1.35	1.10	11.41	2.05
Rheumatology and Immunology	10.18	6.71	5.15	12.30	0.87	0.84	1.85	0.79	9.45	1.30
Dermatology	12.16	7.74	6.84	10.93	1.25	2.15	1.26	1.17	10.75	2.05
Average	11.50	7.68	6.59	10.99	1.09	1.64	1.28	0.92	9.91	1.48

Table 5: NDCG@10 scores for ten retrieval models representing different retrieval styles, including lexicon-based retrievers, uni-modal dense retrievers, CLIP-style models, multi-modal dense retrievers, and late-interaction multi-modal retrievers for the Case study Retrieval task in the *M3Retrieve* Benchmark.

Image Integrator (MII) offer modest yet consistent improvements (averages of 1.09 and 1.64, respectively), with MII peaking in Cardiology (2.31). Vision–language adapters, CLIP SF and BLIP FF, deliver complementary gains: CLIP SF excels in Rheumatology and Immunology (1.85) and Cardiology (2.08), and BLIP FF adds notable lift in Dermatology (1.17). Finally, the unified FLMR model achieves robust performance across domains, reaching its highest scores in Genetics and Genomics (2.45) and Gastroenterology (2.05), underscoring its versatility for image-informed medical retrieval.

A.2.3 Query to Image Retrieval

Domain-wise Model Performance Analysis

Table 6 presents the performance of various models across different anatomical domains, namely Spine and Muscles, Abdomen, Head, Thorax, and the Reproductive and Urinary System.

In the **Spine and Muscles** domain, the highest performance is observed with the MII model (37.88), while CLIP SF and MM Embed also demonstrate competitive performance (28.65 and 26.01, respectively). BLIP FF shows moderate alignment (6.00), while MM Ret and FLMR report lower scores (3.50 and 2.10).

In the **Abdomen**, MII achieves the strongest result (49.04), followed by CLIP SF (29.41) and MM Embed (27.54). Performance from BLIP FF (1.22), MM Ret (1.45), and FLMR (3.00) remains modest.

The **Head** domain reveals comparatively lower performance across most models. MII again leads (34.31), while MM Embed and CLIP SF register similar outcomes (26.54 and 19.38, respectively). Other models exhibit limited effectiveness.

For the **Thorax**, both MII (47.27) and MM Embed (35.03) perform strongly, with CLIP SF also maintaining a good score (33.75). However, MM Ret (3.38), BLIP FF (1.00), and FLMR (2.80) show relatively lower results.

Lastly, in the **Reproductive and Urinary System**, MII (49.14) and CLIP SF (34.13) dominate, while MM Embed also performs reasonably well (32.36). The remaining models perform below par in this category.

Overall, MII consistently achieves the highest average performance (43.53), indicating robust multi-domain capabilities. CLIP SF (29.12) and MM Embed (29.49) also show strong generalization. In contrast, BLIP FF (2.07), MM Ret (2.28), and FLMR (2.56) underperform on average, indicating more limited domain adaptation.

Discipline	MM Ret	MII	CLIP SF	BLIP FF	MM Embed	FLMR
Spine and Muscles	3.50	37.88	28.65	6.00	26.01	2.10
Abdomen	1.45	49.04	29.41	1.22	27.54	3.00
Head	2.05	34.31	19.38	1.69	26.54	2.50
Thorax	3.38	47.27	33.75	1.00	35.03	2.80
Reproductive and Urinary System	1.01	49.14	34.13	1.28	32.36	2.40
Average	2.28	43.53	29.12	2.07	29.49	2.56

Table 6: NDCG@10 scores for ten retrieval models representing different retrieval styles, including lexicon-based retrievers, uni-modal dense retrievers, CLIP-style models, multi-modal dense retrievers, and late-interaction multi-modal retrievers for the Query to Image Retrieval task in the *M3Retrieve* Benchmark.