MedCore: Empowering LVLMs for Medical Tasks with Multimodal Database and RAG

CHAN Shu Pui, LAM Sum Ying, YE Yilin and YIP Sau Lai

HC1

Advised by Prof. CHEN Hao

Abstract

The healthcare sector struggles with physician shortages and extended patient wait times, driving the need for innovative solutions. Large vision-language models (LVLMs) show promise in aiding medical diagnoses and triage, but their general-domain training limits medical expertise, and specialized training is costly. We propose MedCore, a cost-effective system to enhance LVLMs' medical performance without retraining. MedCore integrates MedCore-DB, a multimodal knowledge base and visual question-answering (VQA) benchmark built from medical documents and webpages, and MedCore-RAG, a novel retrieval-augmented generation framework. MedCore-RAG retrieves relevant image-text snippets from MedCore-DB to augment LVLM responses for tasks like VQA and medical report generation. Evaluations demonstrate improved accuracy and relevance, highlighting MedCore's potential to support healthcare applications while addressing scalability and cost challenges.

1 Introduction

The healthcare industry faces significant challenges due to a shortage of human resources, leading to overloaded physicians and prolonged patient wait times. To address these issues, LVLMs are increasingly explored for assisting doctors in diagnoses. However, general-domain LVLMs often lack specialized medical knowledge, and training such models for medical applications is costly. This study proposes ***MedCore***, a cost-effective and flexible assistant system to enhance the medical performance of general LVLMs without additional training comprising two components: ***MedCore-DB*** and ***MedCore-RAG***. MedCore-DB includes a knowledge base constructed from medical documents and websites and a VQA benchmark to evaluate generalist medical LVLMs. MedCore-RAG introduces a novel multimodal retrieval-augmented generation (RAG) framework that retrieves relevant information from the knowledge base to bridge the medical knowledge gap in LVLMs. This framework supports various medical tasks, such as VQA and medical report generation. During inference, users can query the system with medical tasks by inputting images and text. MedCore-RAG then uses MedCore-DB as its corpus to retrieve relevant snippets through similarity search. These snippets are used to generate a prompt, which is fed into a frozen LVLM to produce an augmented response for the user.

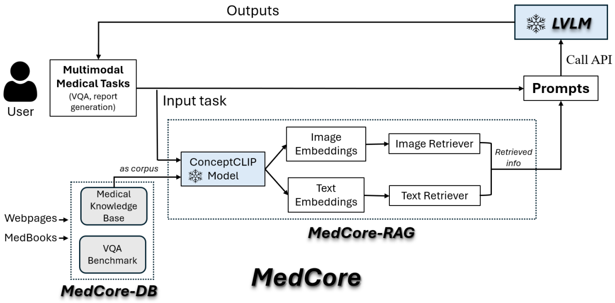


Figure 1 System architecture of ***MedCore****.*

2 Methodology

2.1 MedCore-DB

2.1.1 Overview of *MedCore-DB*

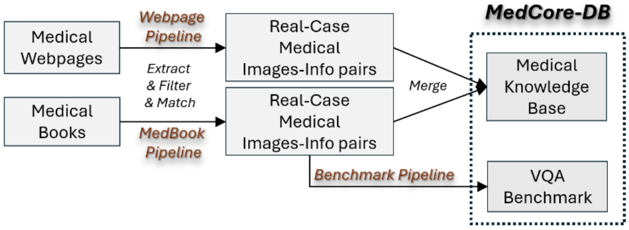


Figure 2 Overview of MedCore-DB

MedCore-DB consists of a knowledge base database and a VQA benchmark to be used as the corpus andtesting set for the whole system. The knowledge database contains medical image-information pairs extracted from open-sourced medical webpages and books using separated them, namely the ***Webpage Pipeline*** and the ***MedBook Pipeline***. VQA are then generated and carefully filtered in the ***Benchmark Pipeline***.

2.1.2 *Webpage Pipeline*: Knowledge Basd from Medical Webpages

The webpage knowledge base consists of two components: an offline corpus and an online real-time search and parsing toolkit. This design ensures that MedCord-DB is equipped with extensive medical knowledge while also keeping information current.

We initially developed the Webpage Pipeline to establish the offline knowledge base, and later utilized it to build a toolkit that can parse webpages given the query keywords.

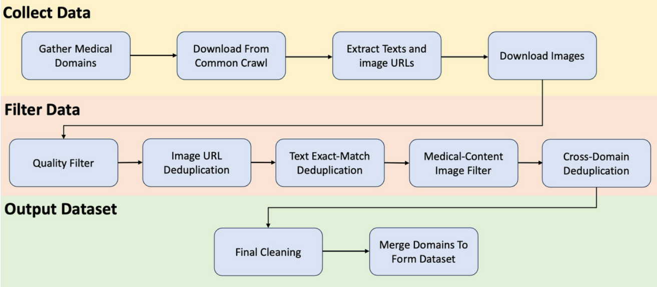

Collect Data
Gather Medical Domains
Download From Common Crawl
Extract Texts and image URLs
Download Images
Filter Data
Quality Filter
Image URL Deduplication
Text Exact-Match Deduplication
Medical-Content Image Filter
Cross-Domain Deduplication
Output Dataset
Final Cleaning
Merge Domains To Form Dataset

Figure 3 Flow chart for Webpage pipeline.

(1) Offline Knowledge Base Stage 1: Collect Medical Related Webpages

CommonCrawl serves as an open-source repository containing billions of webpages, and a subset of our dataset was constructed by extracting medically relevant content from this resource. Since CommonCrawl does not inherently categorize webpages by domain, the initial task involved identifying and retrieving a medical-oriented subset. We used a list of diseases keywords to query google and collect the returned URL to construct a medical related domain list. We then download these websites from CommonCrawl foaming out raw dataset.

(2) Offline Knowledge Base Stage 2: Text and Image Data Extraction

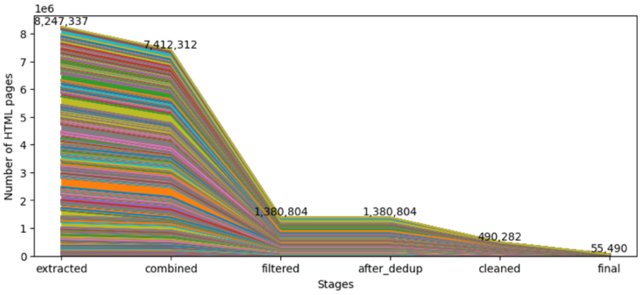
Typical HTML files include scripts and styles that is irrelevant to its content. Additionally, webpages often contain boilerplate content (e.g., navigational elements like "Home," "Go Back," or "View More"). This necessitates multi-level filtering and structural simplification processes. We first implemented a preliminary filtering process to eliminate non-structural and non-content elements, including CSS and JS code. This initial filtering stage also excluded documents lacking embedded image links, aligning with our objective to construct a multimodal dataset.

Subsequently, we applied a simplification algorithm to extract only the main content. Through these foundational preprocessing techniques, we reduced 84% content from the original collection.

(3) Offline Knowledge Base Stage3: Non-Medical & Duplicated Content Filtering

To further refine the data, we executed intra-document and global text-image deduplication processes to eliminate website logos, icons, residual boilerplate content, and redundant textual elements across multiple domains. Finally, leveraging Large Vision-Language Models (LVLMs), we systematically classified each image in the dataset to retain only medically relevant visuals. The resultant curated multimodal medical web dataset comprises 55,490 HTML documents, each containing at least one relevant medical image.

Figure 4 Number of HTMLs at each stage



2.1.3 *MedBook Pipeline*: Knowledge Base from Medical Books

(1) Oveview

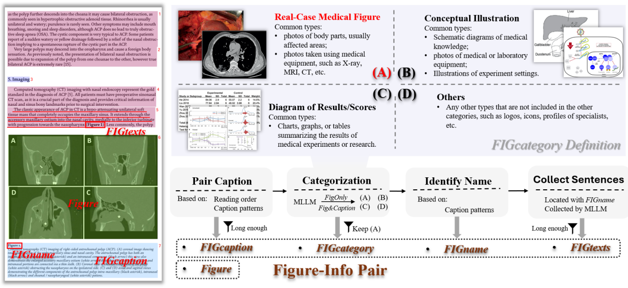

FIGtexts
Figure
FIGname
FIGcaption
Real-Case Medical Figure
Common types:
photos of body parts, usually affected areas;
photos taken using medical equipment, such as X-ray, MRI, CT, etc.
Conceptual Illustration
Common types:
Schematic diagrams of medical knowledge;
photos of medical or laboratory equipment;
Illustrations of experiment settings.
(A)
(B)
(C)
(D)
Diagram of Results/Scores
Common types:
Charts, graphs, or tables summarizing the results of medical experiments or research.
Others
Any other types that are not included in the other categories, such as logos, icons, profiles of specialists, etc.
FIGcategory Definition
Pair Caption
Based on: Reading order Caption patterns
Categorization
MLLM FigOnly (A) (B) Fig&Caption (C) (D)
Identify Name
Based on: Caption patterns
Collect Sentences
Located with FIGname Collected by MLLM
Long enough
Keep (A)
Long enough
FIGcaption
FIGcategory
FIGname
FIGtexts
Figure
Figure-Info Pair

*Figure 5 Overview of **MedBook Pipeline**.*

(2) Stage 1: Book Collection and Preprocessing.

8,090 books were downloaded by querying DOAB with list of medical keywords. Books are then processed using MinerU [1], a layout detection tool that divides each page into smaller regions, labels them by type (e.g., image, text, title) and organizes them in reading order. Although caption pairing is performed, there are captions misclassified as text blocks. Recovery is performed using pattern matching, since captions usually start with reference names like "Figure 1.1" or "Fig 2".

(3) Stage 2: Figure & Information

Relevant information are collected at this stage, including figure captions (***FIGcaption***), figure categories (***FIGcategory***), figure reference names (***FIGname***), and sentences in paragraphs that describe the image by explicitly mentioning the name (***FIGtexts***). The *FIGcaption* is paired according to reading order. Successfully paired images are then categorized into four categories: **(A) real-case medical figure, (B) conceptual illustration, (C) diagram of results/scores, (D) other**, by prompting the InternVL2-8B [2] model with two approaches: (a) FigOnly: inputting the figure only and (b) Fig&Caption: inputting both the figure and its caption. Only images categorized as category (A) in either approach are retained. The *FIGname* is then extracted from the caption using pattern matching rules, which is used as a location identifier to colloect *FIGtexts*, sentences that explicitly mention the *FIGname*, from paragraphs on the previous, current, and subsequent pages by prompting InternVL2-8B [2]. Additionally, images with insufficient textual information (i.e., total words of *FIGcaption* and *FIGtexts* less than 5) are excluded. This stage result in a total of 36,820 pairs.

2.1.4 ***Benchmark Pipeline*: VQA Benchmark from Image-Text Pairs**

The ***Benchmark Pipeline*** starts by generationg five types of VQA pairs using InternVL2.5-78B [2], including: **(S) Symptom Recognition, (SO) Sugery & Operation, (D) Disease Diagnosis, (M) Modality Recognition** and **(A) Anatomy Identification**. The VQAs are reformatted into multiple-choice questions (MCQs) using qwen-vl-max-latest [3]. Due to the concern of computational cost and benchmark size, 10,000 VQAs were randomly selected for further processing. Three steps are then conducted to filter inapproate VQAs. Fistly, a LVLM qwen-vl-max-latest [3] was prompted to determine suitability and information consistency. Secondly, VQAs determined as answerable without the image by a LLM, DeepSeek-R1 [4] are excluded since the image is redudent. Manual verification is perpromed at the end. These steps result in 6971 VQAs. 1,000 are randomly selected for each VQA type to form the final VQA benchmark, 5,000 entries in total.

The benchmark is organized with a hierarchical labeling system by categorizing MCQ based on medical modality, anatomy, and department. The benchmark is diverse and comprehensive, covring 51 modalities, 133 organs and 35 departments at the most granual level.

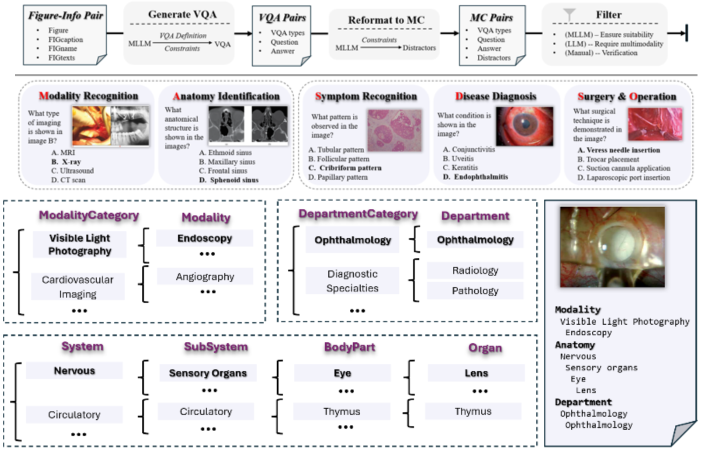


Figure 6 Overview of ***Benchmark Pipeline*** and label system.

2.2 MedCore-RAG

2.2.1 Sequential pipeline

Our MedCore-RAG system leverages pipelines with for two significant multimodel medical tasks, including VQA and radiology report generation. In sequential pipeline, the Visual-Only Retriever first identifies relevant instances through queried images, followed by the Text-Only Retriever, which then processes the retrieved instances to ensure textual data is aligned with the question.

Pre-embedding features and efficient searching The preprocessing component builds upon the principles of the LongRAG framework [5], which minimizes the retriever's workload by grouping related documents into larger units. Inspired by this methodology, we consolidate VQA pairs corresponding to the same medical images into larger data samples, ensuring that relevant information is retained while optimizing retrieval efficiency. For the report generation task, we extract the annotation of each image as its corresponding textual description. Both image and text features are embedded using ConceptCLIP in advance to enable efficient similarity-based retrieval. To further accelerate the search process, we utilize FAISS [6] for fast and scalable similarity search over the embedded features.

Retrieval components The Visual-Only and Text-Only Retrievers work in tandem to search for the most relevant samples from the knowledge base based on their alignment with the query data. These components are inspired by the EchoSight [7] and RAMM [1], with the primary objective being to retrieve the most similar instances—whether visual or textual—to supplement the LVLMs. Both components leverage FAISS [6] to accelerate similarity computations, resulting in a tenfold improvement in retrieval speed.

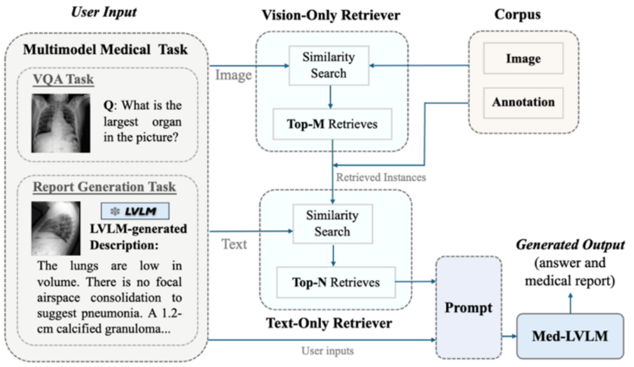
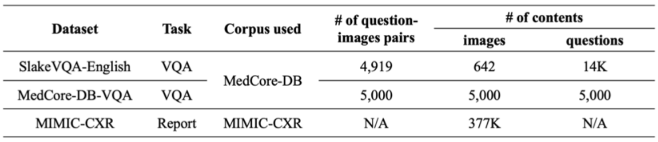


Figure 7. The overall frameworks of MedCore-RAG - Sequential Retrieval: framework incorporates preprocessing, followed by sequential Vision-Only and Text-Only Retrieval.

2.2.2 Experiment setup

During the experimental phase, the proposed MedCore-RAG framework will be evaluated using a variety of VQA datasets and a report generation dataset. During testing, we utilized Qwen2.5-VL-72B-Instruct-AWQ as the backbone model and accessed it via API.

Testing datasets The details of datasets used are listed below:



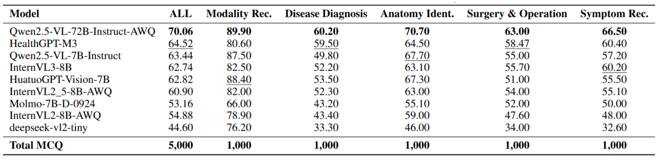
Dataset	Task	Corpus used	# of question-images pairs	# of contents	
				images	questions
SlakeVQA-English	VQA	MedCore-DB	4,919	642	14K
MedCore-DB-VQA	VQA		5,000	5,000	5,000
MIMIC-CXR	Report	MIMIC-CXR	N/A	377K	N/A

Figure 8 Table of datasets used.

Performance metrics For VQA task, the queries in the evaluation dataset can be categorized into three types: multiple choice, close questions and open questions. We use accuracy for multiple choice question and close questions, and F1 score for open questions. For report generation task, we use METEOR, BLEU-1, BLUE-4, F1 score and precision as metrics.

3 Results

For MedCore-DB, the benchmark was tested on several popular general LVLMs to assess the effectiveness of the benchmark and the knowledge base, including QwenVL2.5 (8B and 72B) [3], InternVL2-8B, InernVL2.5-8B and InternVL3-8B [2], Molmo-7B-D-0924 [8], and DeepSeekVL2-tiny [9], and a list of medical LVLMs including HealthGPT-M3 [10], HuatuoGPT-Vison-7B [11]. Each LVLM is prompt to answer all the MCs in the benchmark, with temperature set as 0. Accurary for all and each type of VQA is determined with the percentage of corrected answered MCs.

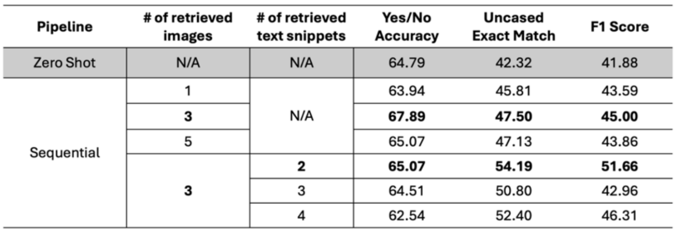


Model	ALL	Modality Rec.	Disease Diagnosis	Anatomy Ident.	Surgery & Operation	Symptom Rec.
Qwen2.5-VL-72B-Instruct-AWQ	**70.06**	**89.90**	**60.20**	**70.70**	**63.00**	**66.50**
HealthGPT-M3	64.52	80.60	59.50	64.50	58.47	60.40
Qwen2.5-VL-7B-Instruct	63.44	87.50	49.80	67.70	55.00	57.20
InternVL3-8B	62.74	82.50	52.20	63.10	55.70	60.20
HuatuoGPT-Vision-7B	62.82	88.40	53.50	67.30	51.00	55.50
InternVL2_5-8B-AWQ	60.90	82.00	52.30	63.00	54.00	55.10
Molmo-7B-D-0924	53.16	66.00	43.20	55.10	52.00	50.00
InternVL2-8B-AWQ	54.88	78.90	43.40	59.00	47.60	48.00
deepseek-vl2-tiny	44.60	76.20	33.30	46.00	34.00	32.60
Total MCQ	**5,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**

Figure 9 Evaluation results on the VQA benchmark. The scores are the percentages of correctly answered MCs.

For MedCore-RAG, three experiments were conducted to test the performance of the pipeline for two tasks.

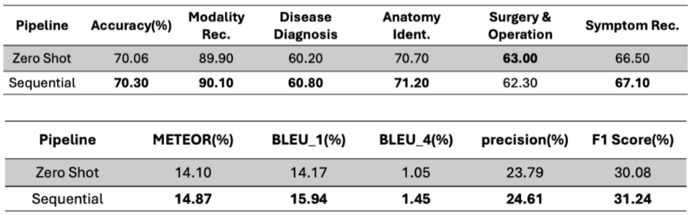
Parameters tuning The following are the results of our experiments on SlakeVQA. By first testing on the number of retrieved images and then on the number of retrieved text N, we have the best parameter setting of (3,2), which means we will take three images from the corpus by image similarities and only two of them with higher text similarities will be prompted in LVLMs in future experiments across various datasets.



Pipeline	# of retrieved images	# of retrieved text snippets	Yes/No Accuracy	Uncased Exact Match	F1 Score
Zero Shot	N/A	N/A	64.79	42.32	41.88
Sequential	1	N/A	63.94	45.81	43.59
	3		**67.89**	**47.50**	**45.00**
	5		65.07	47.13	43.86
	3	**2**	**65.07**	**54.19**	**51.66**
		3	64.51	50.80	42.96
		4	62.54	52.40	46.31

Figure 10 Results of parameter tuning on SlakeVQA.

Tasks We test the performance for VQA task on MedCore-DB and for radiology report generation task on MIMIC-CXR. The results have shown that sequential pipeline has outperformed zero shot by about 4%.



Pipeline	Accuracy(%)	Modality Rec.	Disease Diagnosis	Anatomy Ident.	Surgery & Operation	Symptom Rec.
Zero Shot	70.06	89.90	60.20	70.70	**63.00**	66.50
Sequential	**70.30**	**90.10**	**60.80**	**71.20**	62.30	**67.10**

Pipeline	METEOR(%)	BLEU_1(%)	BLEU_4(%)	precision(%)	F1 Score(%)
Zero Shot	14.10	14.17	1.05	23.79	30.08
Sequential	**14.87**	**15.94**	**1.45**	**24.61**	**31.24**

Figure 11 Results of VQA task on MedCore-DB (up) and results of medical report generation task on MIMIC-CXR (down).

4 Conclusion

The MedCore system, comprising MedCore-DB and MedCore-RAG, successfully advances the medical capabilities of general large vision-language models (LVLMs). MedCore-DB achieved its objectives by developing effective pipelines to construct a comprehensive medical knowledge base and a visual question-answering (VQA) benchmark from medical webpages and books. Tailored processing steps, including medical-content filters and scope restrictions, enabled the use of general tools for domain-specific tasks, with the database's diversity validated through data distribution and model performance analyses.

Future work for MedCore-DB includes enhancing the Webdata pipeline with dedicated parsing for high-value domains (e.g., ultrasound case studies), improving the MedBook pipeline’s adaptability across diverse book types, and training lightweight classifiers for data labeling to boost efficiency. For MedCore-RAG, planned improvements involve integrating a reranker to refine retrieval accuracy, incorporating domain-specific features for radiology and pathology (e.g., cell graphs), and benchmarking against systems like RULE to guide optimizations. These efforts aim to enhance MedCore’s accuracy, scalability, and clinical relevance, paving the way for broader real-world impact.

References

[1] W. Bin, X. Chao, Z. xiaomeng, O. Linke, W. Fan, Z. Zhiyuan, X. Rui, L. Kaiwen, Q. Yuan, S. Fukai and et.al, “MinerU: An Open-Source Solution for Precise Document Content Extraction,” *arxiv.*

[2] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu and L. Lu, “Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks,” arXiv preprint arXiv:2312.14238, 2023.

[3] Y. An and et.al, “Qwen2.5 Technical Report,” *arXiv.*

[4] DeepSeek-AI, *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning,* https://arxiv.org/abs/2501.12948, 2025.

[5] Z. Jiang, X. Ma and W. Chen, “LongRAG: Enhancing Retrieval-Augmented Generation with Long-context LLMs,” arxiv preprint arXiv:2406.15319, 2024.

[6] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini and H. Jégou, The Faiss library, arXiv2401.08281, 2024.

[7] S. Sarto, M. Cornia, L. Baraldi and R. Cucchiara, “Retrieval-augmented transformer for image captioning,” in *International Conference on Content-based Multimedia Indexing*, 2022.

[8] e. a. Matt Deitke, *Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models,* https://arxiv.org/abs/2409.17146, 2024.

[9] e. Zhiyu Wu, *DeepSeek-VL2: Mixture-of-Experts Vision-Language Models for Advanced Multimodal Understanding,* https://arxiv.org/abs/2412.10302, 2024.

[10] e. Tianwen Lin, *HealthGPT: A Medical Large Vision-Language Model for Unifying Comprehension and Generation via Heterogeneous Knowledge Adaptation,* https://arxiv.org/abs/2502.09838, 2025.

[11] J. Chen and et.al, “HuatuoGPT-Vision, Towards Injecting Medical Visual Knowledge into Multimodal LLMs at Scale,” arXiv, 6 2024. [Online]. Available: https://arxiv.org/abs/2406.19280.

[12] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman and S. Anadkat, “Gpt-4 technical report,” arXiv preprint arXiv:2303.08774, 2023.

[13] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong and T. Yu, “Palm-e: An embodied multimodal language model,” in *International Conference on Machine Learning*, 2023.

[14] H. Liu, C. Li, Q. Wu and Y. J. Lee, “Visual instruction tuning,” *Advances in neural information processing systems,* vol. 36, 2024.

[15] K. Guu, K. Lee, Z. Tung, P. Pasupat and M. Chang, “Retrieval augmented language model pretraining,” in *International Conference on Machine Learning,*, 2020.

[16] K. Lee, M. Chang and K. Toutanova, “Latent retrieval for weakly supervised open domain question answering,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, July, 2019.

[17] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kuttler, M. Lewis, W.-t. Yih and T. Rocktaschel, “Retrieval-augmented generation for knowledge-intensive nlp tasks,” *Advances in Neural Information Processing Systems,* vol. 33, pp. 9459-9474, 2020.

[18] A. Long, W. Yin, T. Ajanthan, V. Nguyen, P. Purkait, R. Garg, A. Blair, C. Shen and A. v. d. Hengel, “Retrieval augmented classification for long-tail visual recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[19] Y. Yan and W. Xie, “EchoSight: Advancing Visual-Language Models with Wiki Knowledge,” arXiv preprint arXiv:2407.12735v1, 2024.

[20] P. Xia, K. Zhu, H. Li, H. Zhu, Y. Li, G. Li, L. Zhang and H. Yao, “RULE: Reliable Multimodal RAG for Factuality in Medical Vision Language Models,” arxiv preprint arXiv:2407.05131v1, 2024.

[21] Z. Yuan, “Minigpt-4: Enhancing vision-language understanding with advanced large language models,” in *The Twelfth International Conference on Learning Representations*, 2023.

[22] Z. Yuan, Q. Jin, C. Tan, Z. Zhao, H. Yuan, F. Huang and S. Huang, “RAMM: Retrieval-augmented Biomedical Visual Question Answering with Multi-modal Pre-training,” *Proceedings of the 31st ACM International Conference on Multimedia,* p. 547–556, 2023.

[23] B. Liu, L.-M. Zhan, L. Xu, L. Ma, Y. Yang and X.-M. Wu, “Slake: A semanticallylabeled knowledge-enhanced dataset for medical visual question answering,” in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 2021.

[24] X. He, Y. Zhang, L. Mou, E. Xing and P. Xie, “Pathvqa: 30000+ questions for medical visual question answering,” arXiv preprint arXiv:2003.10286, 2020.

[25] Y. Hu, T. Li, Q. Lu, W. Shao, J. He, Y. Qiao and P. Luo, “OmniMedVQA: A New Large-Scale Comprehensive

Evaluation Benchmark for Medical LVLM," arxiv preprint arXiv:2402.09181, 2024.

[26] P. Chen, J. Ye, G. Wang, Y. Li, Z. Deng, W. Li, T. Li, H. Duan, Z. Huang, Y. Su, B. Wang, S. Zhang, B. Fu, J. Cai, B. Zhuang, E. J. Seibel, J. HE and Y. Qiao, "GMAI-MMBench: A Comprehensive Multimodal Evaluation Benchmark Towards General Medical AI," arxiv preprint arXiv:2408.03361, 2024.

[27] H. Laurençon and e. al., "OBELICS: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents," *arXiv preprint,* vol. 2306.16527v2, 2023.

[28] e. a. Renqiu Xia, "DocGenome: An Open Large-scale Scientific Document Benchmark for Training and Testing Multi-modal Large Language Models," arXiv, 6 2024. [Online]. Available: https://arxiv.org/abs/2406.11633.

[29] A. Awadalla and e. al, "MINT-1T: Scaling Open-Source Multimodal Data by 10x: A Multimodal Dataset with One Trillion Tokens," arXiv, 6 2024. [Online]. Available: https://arxiv.org/abs/2406.11271.

[30] Y. Hu and et.al, "OmniMedVQA: A New Large-Scale Comprehensive Evaluation Benchmark for Medical LVLM," arXiv, 2 2024. [Online]. Available: https://arxiv.org/abs/2402.09181.

[31] X. Zhang and et.al, "PMC-VQA: Visual Instruction Tuning for Medical Visual Question Answering," arXiv, 5 2023. [Online]. Available: https://arxiv.org/abs/2305.10415.

[32] e. Yunfei Xie, "MedTrinity-25M: A Large-scale Multimodal Dataset with Multigranular Annotations for Medicine," arXiv, 8 2024. [Online]. Available: https://www.arxiv.org/abs/2408.02900.

[33] CommonCrawl, "Common Crawl," [Online]. Available: https://commoncrawl.org.

[34] D. Müller, I. Soto-Rey and F. Kramer, "TOWARDS A GUIDELINE FOR EVALUATION METRICS IN MEDICAL IMAGE SEGMENTATION," arXiv, 2 2022. [Online]. Available: https://arxiv.org/pdf/2202.05273.

[35] M. Moor and et.al, "Med-Flamingo: a Multimodal Medical Few-shot Learner," arXiv, 7 2023. [Online]. Available: https://arxiv.org/abs/2307.15189.

[36] X. Zhao and et.al, "PDF-Extract-Kit," github, 2024. [Online]. Available: https://github.com/opendatalab/PDF-Extract-Kit.

[37] R. Kittinaradorn and et.al, "EasyOCR," github, 2020. [Online]. Available: https://github.com/JaidedAI/EasyOCR.

[38] ATF, "AfraTafreeh | Free content for medical students," [Online]. Available: https://afratafreeh.com/.

[39] Seongsu Bae , Daeun Kyung , Jaehee Ryu , Eunbyeol Cho , Gyubok Lee , Sunjun Kweon , Jungwoo Oh , Lei JI , Eric Chang , Tackeun Kim , Edward Choi, "MIMIC-Ext-MIMIC-CXR-VQA: A Complex, Diverse, And Large-Scale Visual Question Answering Dataset for Chest X-ray Images," 9 July 2024. [Online]. Available: https://physionet.org/content/mimic-ext-mimic-cxr-vqa/1.0.0/. [Accessed 13 Nov 2024].

[40] Alistair Johnson , Tom Pollard , Roger Mark , Seth Berkowitz , Steven Horng, "MIMIC-CXR Database (version 2.1.0)," 23 July 2024. [Online]. Available: https://doi.org/10.13026/4jqj-jw95. [Accessed 13 Nov 2024].

[41] "Directory of Open Access Books," [Online]. Available: https://www.doabooks.org/.

[42] "Directory of Open Access Books," [Online]. Available: https://www.doabooks.org/.

[43] Johnson, Jeff and Douze, Matthijs and Jegou, Herve, "Billion-scale similarity search with GPUs," *IEEE Transactions on Big Data,* vol. 7, pp. 535--547, 2019.

[44] e. S. K. M. S. Islam, *Introduction of Medical Imaging Modalities,* https://arxiv.org/abs/2306.01022, 2023.

[45] wikipedia, "Meidcal imaging," [Online]. Available: https://en.wikipedia.org/wiki/Medical_imaging. [Accessed 11 2024].

[46] wikipedia, "List of organs of the human body," [Online]. Available: https://en.wikipedia.org/wiki/List_of_organs_of_the_human_body. [Accessed 11 2024].

[47] wikipeida, "List of skeletal muscles of the human body," [Online]. Available: https://en.wikipedia.org/wiki/List_of_skeletal_muscles_of_the_human_body. [Accessed 11 2024].

[48] wikipedia, "Tendon," [Online]. Available: https://en.wikipedia.org/wiki/Tendon#List_of_Tendons. [Accessed 11 2024].

[49] wikipeidia, "List of bones of the human skeleton," [Online]. Available: https://en.wikipedia.org/wiki/List_of_bones_of_the_human_skeleton. [Accessed 11 2024].

[50] M. CLINIC, "Medical Departments and Centers," [Online]. Available: https://www.mayoclinic.org/departments-centers. [Accessed 03 2025].

[51] Y. Nie, S. He, Y. Bie, Y. Wang, Z. Chen, S. Yang and H. Chen, ConceptCLIP: Towards Trustworthy Medical AI via Concept-Enhanced Contrastive Langauge-Image Pre-training, arXiv:2501.15579, 2025.