Multi-Agent Collaboration for Integrating Echocardiography Expertise in Multi-Modal Large Language Models

Yi Qin, Dinusara Sasindu GAMAGE NANAYAKKARA, and Xiaomeng Li✉

Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, China
eexmli@ust.hk

Abstract. Current echocardiography MLLMs rely on diagnostic-focused data lacking detailed image-text descriptions and systematic multi-modal cardiac knowledge, resulting in suboptimal performance across diverse echocardiography visual question answering tasks. Existing methods to integrate clinical expertise face three key challenges when adapting to echocardiography: labor-intensive curation processes, overlooking textual or diagrammatic knowledge sources essential in cardiac diagnosis, and incompatibility with pretrained MLLMs. To address these gaps, we propose Multi-Agent Collaborative Expertise Extractor (MACEE), a multi-agent framework employing MLLM-powered agents to extract echocardiography expertise from diverse sources. MACEE collects the EchoCardiography Expertise Database (ECED), the first comprehensive knowledge repository covering 100+ common and rare cardiac conditions from textbooks, guidelines, and case studies. To integrate ECED into MLLMs, we introduce Echocardiography Expertise-enhanced Visual Instruction Tuning (EEVIT), a lightweight training framework using expertise-guided instruction tuning. EEVIT employs adapters in vision and language modules, enabling efficient expertise integration while training less than 1% of the model's parameters. Experiments validates the effectiveness of these three components. Codes and license details: https://github.com/xmed-lab/ECED

Keywords: Echocardiography · Knowledge Database · Multi-modal Large Language Model.

1 Introduction

Echocardiography is a widely used cardiac imaging modality [18], providing critical diagnostic insights for various cardiac diseases [22,24,23]. Despite recent advancements in echocardiography vision-language models [4,15] and medical multi-modal large language models (MLLMs) [2,17,21], their performance on diverse echocardiography-related visual question answering tasks remains suboptimal (see Table 1). These models are often trained on data collected from hospitals for diagnostic purposes, which usually lack crucial background information, such as the correlation between echocardiography findings and disease

severity (see Fig. 1 (a)). This misaligns the model's learning and decision process with that of clinicians [12], leading to a lack of essential medical knowledge required for interpreting multi-modal inputs and unreliable predictions. Furthermore, no comprehensive echocardiography knowledge database currently exists to address this gap. Therefore, it is essential to inject echocardiography clinical knowledge for further echocardiography analysis performance enhancement in MLLMs.

Recently, several methods [20,25,19,26,29,28] have leveraged clinical knowledge from medical literature during the training process of vision-language models to enhance model performance more efficiently. However, three key limitations constrain their ability to inject rich echocardiographic expertise into MLLMs. *First, most expertise curation methods [20,11,19,26] require manual effort for aligning image-text pairs, matching keywords, and cleaning raw contents.* They rely on basic rule-based extraction techniques and additional algorithms for image processing, requiring extensive manual designs for each form of knowledge source (e.g., six people working for four weeks [20]). Given the diverse guidelines and inspection modalities in echocardiography, such manual approaches are insufficient for building a comprehensive expertise database. *Second, current methods focus mostly on collecting image-caption pairs.* However, effective diagnosis in echocardiography often requires multiple sources of evidence and reasoning steps, a type of knowledge that would usually present in pure text, tables, and diagrams in textbooks and guidelines (see Fig. 1 (c) for examples). These rich sources of diagnostic knowledge remain largely underutilized. *Third, existing approaches primarily target training discriminative vision-language models from scratch,* neglecting strategies to inject expertise into pretrained MLLMs. As a result, effectively leveraging the extensive visual comprehension abilities of pretrained MLLMs to integrate medical expertise and enhance performance remains an open challenge. More flexible and scalable curation methods are needed to accommodate diverse types of expertise and fully exploit the potential of pretrained MLLMs in echocardiographic tasks.

To address these challenges, our core innovation is to streamline the expertise curation process by employing multiple MLLM-powered agents to compre-

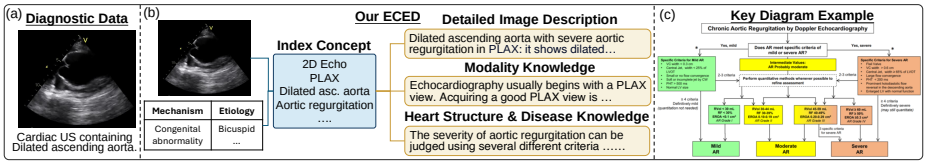

Fig. 1. The comparison between existing MLLM training data source and our proposed EchoCardiography Expertise Database (ECED). (a) The label of existing dataset only contains brief, diagnostic conclusion. (b) Our ECED provides detailed image descriptions along with comprehensive echocardiography background knowledge in diverse formats, which readily support MLLM performance enhancement. (c) An example of a key diagram essential in assessing Aortic Regurgitation [30].

hensively collect echocardiographic knowledge from diverse knowledge sources. Specifically, we propose the **M**ulti-**A**gent **C**ollaborative **E**xpertise **E**xtractor (MACEE), a novel framework comprising five MLLM-based agents collaboratively accelerate the expertise extraction pipeline. MACEE leverages state-of-the-art document and text-processing capabilities of MLLMs to extract content, align images, clean textual descriptions, and summarize contents with high efficiency. The result of MACEE is the **E**cho**C**ardiography **E**xpertise **D**atabase (ECED), the first comprehensive echocardiography knowledge database. The ECED is curated from four authoritative textbooks, 44 clinical guidelines from both China and the United States, and a collection of case studies. It captures a wide spectrum of fine-grained knowledge, covering over 100 common and rare heart conditions as well as more than five types of echocardiography modalities, offering an comprehensive repository of echocardiographic expertise to support MLLMs, as Fig. 1 (b) shows. To integrate this curated expertise into MLLMs, we further propose the **E**chocardiography **E**xpertise-enhanced **V**isual **I**nstruction **T**uning (EEVIT) framework. EEVIT organizes diverse knowledge from ECED to augment the training data, ensuring effective expertise injection in model fine-tuning. It utilizes lightweight adapters in both the vision encoder and language decoder of the MLLM, allowing the model to efficiently incorporate and leverage rich echocardiographic expertise while training less than 1% of MLLM's original parameters. We highlight our contributions as follows:

- We propose the Multi-Agent Collaborative Expertise Extractor (MACEE), a novel framework of five MLLM-based agents designed to efficiently organize expertise from diverse sources with less human intervention.
- Based on MACEE, we construct the EchoCardiography Expertise Database (ECED), the first comprehensive echocardiography knowledge database, encompassing over 100 heart conditions and five echocardiography modalities.
- To utilize ECED, we develop Echocardiography Expertise-enhanced Visual Instruction Tuning (EEVIT), which structures diverse expertise from ECED into augmented training data and employs lightweight adapters to integrate echocardiographic knowledge into MLLMs efficiently.
- Experimental results on three public benchmarks demonstrate that EEVIT, powered by ECED, achieves state-of-the-art performance in echocardiography comprehension tasks.

2 The Echocardiography Expertise Database

2.1 Multi-Agent Collaborative Expertise Extractor

The Goal of Expertise Extraction and Pipeline Overview. The objective of MACEE is to construct a concept-indexed database. Each entry in this database comprises two components: **Index Concepts**, which are keywords related to echocardiography and cardiology, summarized from the corresponding content; and **Content**, which can be tables, aligned image-caption pairs, or paragraph segments. Users can efficiently retrieve content from the database by

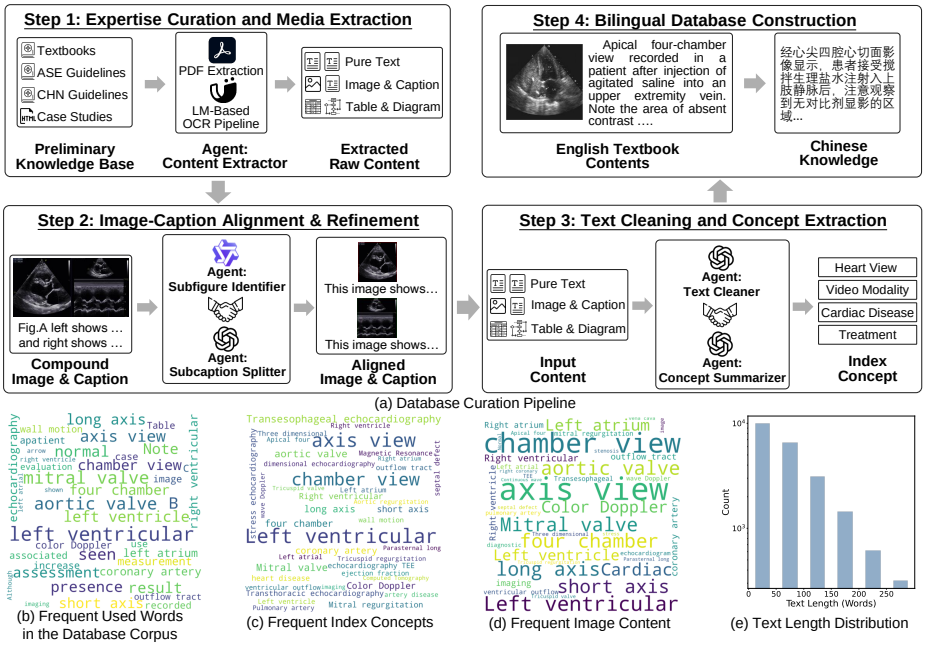

Fig. 2. The pipeline of MACEE and the dataset statistics. (a) The data curation pipeline. (b) Top 40 frequent words in ECED's textual corpus. (c) Top 40 frequent index concepts. (d) Top 40 frequent image content calculated based on the corresponding index concepts. (e) Text length distribution of ECED.

searching for the most relevant index concepts. The overall pipeline of MACEE is depicted in Fig. 2 (a). We followed the agent design principle to create five agents, each for a specific task. Fewer agents (multi-tasking) reduce task-specific performance, while more agents increase unnecessary token costs. Each agent in MACEE is a state-of-the-art MLLM, powered by either GPT-4o [8] or Qwen 2.5-VL [1], and is equipped with our specially designed system prompt amplifying its strength for specific tasks. Notably, the database construction took only two people two weeks to complete. Two board-certified echocardiologists further validated the construction process by confirming the knowledge source is authoritative and comprehensive as well as verifying the key intermediate outputs and final results is clinically correct and relevant.

Step 1: Content Extractor Agent for Expertise Curation and Media Extraction. The Content Extractor agent, combining a professional PDF processor and a state-of-the-art foundation model-based document processor [16], is designed to extract raw content. It processes curated PDF textbooks and web-based HTML files to extract all possible content and categorize the output by type (e.g., pure text, image-caption pair candidates, tables, and diagrams).

Step 2: Subfigure Identifier and Subcaption Splitter Agents Collaboration for Image-Caption Alignment and Refinement. Many extracted

images are compound figures with multiple subfigures, causing ambiguity in caption descriptions. To address this, we propose two agents: Subfigure Identifier and Subcaption Splitter. Instead of training an object detector from scratch [11,14], the Subfigure Identifier detects subfigure bounding boxes directly as the output coordinates of the agent. The Subcaption Splitter collaborates with the Subfigure Identifier to assign corresponding subcaption to each subfigure from the original caption. The output of this step is aligned image-caption pairs.

Step 3: Text Cleaner and Concept Summarizer Agent Collaboration for Text Cleaning and Concept Extraction. The raw content extracted in Step 1 may contain irrelevant information or corrupted characters. To address this, we introduce the Text Cleaner agent to process and organize texts and captions into clean, concise paragraphs suitable for retrieval. Additionally, the Concept Summarizer agent is employed to extract keywords from texts and images, providing a unified interface for capturing multi-modal knowledge. Unlike text-only tools such as MedCAT [9], the Concept Summarizer incorporates information from texts, tables, and images to summarize index concepts more aligned with the data. The agent flexibly summarizes five to ten index concepts based on content length, reducing the risk of hallucinations that may arise from enforcing a fixed number of output. The output of this step includes clean texts, image-caption pairs, tables, and diagrams with corresponding index concepts.

Step 4: Bilingual Database Construction. To facilitate comprehension in both languages, we translate the content from English textbooks into Chinese, creating a bilingual echocardiography knowledge base. However, English guidelines are not translated to Chinese to avoid potential contradictions arising from differences in guideline recommendations between the two countries.

2.2 EchoCardiography Expertise Database Statistics

The current version of ECED incorporates knowledge from four textbooks, 35 ASE guidelines, 9 Chinese guidelines, and 176 online case studies. It includes 4,594 image-caption pairs, 4,233 paragraphs of echocardiography knowledge, and 953 tables and diagrams. All textbook content is available in both English and Chinese, with a vocabulary size of 93,502 words. ECED spans over five modalities, 30 cardiac views, and 100 heart conditions, covering both common conditions like mitral regurgitation and rare cases such as four aortic valve leaflets.

Frequently Used Words. ECED encompasses comprehensive knowledge of echocardiography views, modalities, and heart conditions, forming a rich corpus. Fig. 2 (b) shows the top 40 most frequent words, which include key terms from echocardiography view taxonomy (e.g., axis view, chamber view, short axis) and heart structures (e.g., mitral valve, left ventricular, coronary artery). This word frequency aligns with common views and modalities in clinical practice [13], highlighting ECED's comprehensiveness and clinical relevance.

Frequent Index Concepts. The statistics of index concepts, a key element of ECED, are detailed in Fig. 2 (c), which highlights the 40 most frequent index concepts. These frequent concepts align closely with the word frequency

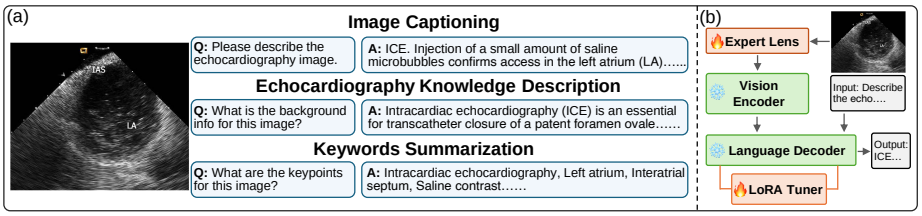

Fig. 3. The training data and model architecture of EEVIT. (a) Examples of augmented training data using ECED. (b) The model architecture of MLLM enhanced by Expert-Lens.

of the entire corpus, making them representative of ECED and facilitating efficient retrieval. Notably, echocardiography modality names, heart structures, and diseases frequently appear among the index concepts, demonstrating that they encompass diverse aspects of knowledge in ECED, thereby ensuring comprehensive retrieval.

Frequent Image Content. ECED includes a comprehensive set of high-quality images sourced from representative figures in textbooks. As shown in Fig. 2(d), these images cover diverse heart structures, echocardiography views, and modalities, enhancing the database's diversity and its coverage of various diseases.

Text Length. Fig. 2 (e) shows the distribution of text lengths in ECED. A total of 44.85% of textual descriptions range from 1 to 50 words, 43.29% range from 50 to 150 words, and 11.85% exceed 150 words. The inclusion of longer, detailed descriptions enhances the richness of image-text pairs and provides abundant textual knowledge.

3 ECED-Augmented Visual Instruction Tuning

Instruction Tuning Data Augmentation using ECED. First, we enhance the instruction tuning dataset by unifying expertise across various formats using index concepts. Specifically, for an ECED entry whose content containing an image, we match its index concepts with those of entries featuring detailed echocardiography-related texts to retrieve these enriched texts. FAISS [5] is employed for the matching process to extract the top-k most relevant entries as background knowledge for the image. In practice, we retrieve the top-2 relevant entries. These retrieved data are then used to construct various Visual Quesition Answering tasks, including image captioning, echocardiography knowledge description, and keyword summarization, as shown in Fig. 3 (a). This method encourages the model to associate echocardiography images with both their contextual and cardiology background knowledge, providing richer information than only performing image captioning task.

Expert-Lens Adapter Enhanced Visual Instruction Tuning. We then perform Expert-Lens Adapter Enhanced Visual Instruction Tuning using the

Table 1. Comparison results with state-of-the-art general and medical MLLMs on three public benchmarks. ↑: the higher, the better results.

Dataset	PMC-VQA	PMC-OA				ROCO-V2			
Method \ Metric	Accuracy ↑	BLEU-1 ↑	F1 ↑	Recall ↑	Bert Score↑	BLEU-1 ↑	F1 ↑	Recall ↑	Bert Score↑
General MLLMs									
InternVL2.5 [3]	0.4273	0.0329	0.0954	0.0712	0.4970	0.0907	0.1316	0.1429	0.5615
Qwen2.5-VL [1]	0.4388	0.0425	0.1293	0.0969	0.5154	0.1460	0.1910	0.2008	0.5874
Medical MLLMs									
MedDr [6]	0.3571	0.0437	0.1060	0.0962	0.5139	0.0729	0.1407	0.1214	0.6010
MedRegA [17]	0.4285	0.0482	0.1176	0.0583	0.5323	0.0630	0.1056	0.1404	0.6131
HuatuoGPT-Vision [2]	0.4532	0.1091	0.1274	0.2186	0.5402	0.0819	0.0864	**0.3178**	0.5613
Ours	**0.5035**	**0.1156**	**0.1676**	**0.2188**	**0.5634**	**0.1610**	**0.1964**	0.2578	**0.6618**

data constructed in the first step, as shown in Fig. 3 (b). Inspired by ViT-Lens [10], we introduce a lightweight self-attention-based adapter, Expert-Lens, at the top of the MLLM vision encoder to specialize it for echocardiography images and videos without fine-tuning all parameters. Expert-Lens enhances the vision encoder's ability to perceive echocardiography features while preserving its capacity to process diverse visual inputs, such as tables and diagrams, enabling effective learning of all formats in ECED. During training, only the parameters of the Expert-Lens and the LoRA [7] adapter in language decoder are updated.

4 Experiments

Dataset and Preprocessing. Following prior works [17,2], we used three benchmarks, including PMC-VQA [27], PMC-OA [11], and ROCO-V2 [14], to evaluate model performance. For all benchmarks, the official test splits were used, and echocardiography-related questions were filtered using the search terms: "echocardiogram OR echocardiography OR cardiac ultrasound OR echocardio OR (cardiac AND ultrasound)." This resulted in 138, 1563, and 360 entries in PMC-VQA, PMC-OA, and ROCO-V2, respectively. **The benchmark data were excluded from development to ensure fairness.** Images were kept at their original resolution and processed using the official code. Training and evaluation procedures followed [17] to ensure reproducibility and fair comparison.

Implementation Details. The method was implemented in PyTorch (v2.1.0) using Qwen2.5-VL-7B as the base model. The Expert-Lens adapter, a two-layer self-attention transformer, was initialized with the top two layer weights of the Qwen2.5-VL vision encoder, following [10]. Training and validation were conducted on four 64GiB MX-C100 TPUs with a batch size $\mathbb{B} = 1$, a learning rate of 1×10^{-4}, and one epoch. Default settings were used for generating predictions for all baselines.

Table 2. The ablation study on ECED's diverse knowledge formats and EEVIT's data augmentation strategy. "-" indicates that the corresponding type of data is excluded from training. "EE" represents the ECED-based instruction data augmentation, while "D" refers to directly adding pure text data as single-modal VQA pairs during training. "T&D" refers to Tables and Diagrams.

Dataset			PMC-VQA	PMC-OA				ROCO-V2			
Image-Text Pairs	T&Ds	Pure Texts	Accuracy ↑	BLEU-1 ↑	F1 ↑	Recall ↑	Bert Score ↑	BLEU-1 ↑	F1 ↑	Recall↑	Bert Score ↑
✓	-	-	0.4748	0.0725	0.1452	0.1256	0.5474	0.1466	0.1735	0.2386	0.6215
✓	✓	-	0.4820	0.0886	0.1574	0.1488	0.5597	0.1521	0.1813	0.2375	0.6348
✓	✓	D	0.4831	0.0925	0.1523	0.1973	0.5611	0.1591	0.1861	0.2542	0.6468
✓	✓	EE	**0.5035**	**0.1156**	**0.1676**	**0.2188**	**0.5634**	**0.1610**	**0.1964**	**0.2578**	**0.6618**

Evaluation Metrics. We followed the existing method [17,2] to report metrics according to task type. For multi-class visual question answering (VQA) (PMC-VQA), we report the test accuracy. For open-ended VQA (PMC-OA and ROCO-V2), we adopt Natural Language Generation (NLG) metrics for performance measurement, including BLEU-1, F1 score, Recall, and BertScore.

Compared with the State-of-the-art MLLMs. Table 1 presents our results compared to state-of-the-art open-source MLLMs, including Qwen2.5-VL [1], InternVL2.5 [3] (general MLLM), and MedRegA [17], MedDr [6], and HuatuoGPT-Vision [2] (medical MLLM). The results reveal that current MLLMs underrepresent echocardiography-related knowledge, resulting in suboptimal performance. Leveraging the expertise from ECED and the effective design of EEVIT, our method outperformed all SOTA MLLMs, significantly improving echocardiographic question answering (PMC-VQA) and content description (PMC-OA and ROCO-V2) performance with only 0.7% trainable parameters. This highlights the importance of integrating high-quality domain expertise to enhance MLLM performance in echocardiography.

Ablation Study. To evaluate the impact of enriching diverse knowledge forms in ECED and the ECED-based data augmentation in EEVIT, we conducted ablation studies to measure the performance gains from different data compositions, as shown in Table 2. Adding high-quality image-text pairs with detailed captions improved overall performance by 2%, highlighting the importance of multi-modal expertise in visual instruction tuning. Further incorporating knowledge in tables, diagrams, and pure text resulted in an additional 7.4% overall performance boost, demonstrating the value of diverse echocardiography expertise. Notably, our ECED-based instruction data augmentation scheme outperformed directly using raw text and tables, confirming the effectiveness of linking echocardiography images to enriched background knowledge through EEVIT.

5 Conclusion

We proposed the Multi-Agent Collaborative Expertise Extractor (MACEE), a novel framework leveraging five MLLM-based agents to efficiently extract ex-

pertise from diverse knowledge sources with less human intervention. Using MACEE, we developed the EchoCardiography Expertise Database (ECED), the first database covering over 100 heart conditions in diverse formats, readily support MLLM performance enhancement. Additionally, we introduced Echocardiography Expertise-enhanced Visual Instruction Tuning (EEVIT), which effectively integrates ECED expertise into pretrained MLLMs by augmenting image-centric instruction data and incorporating Expert-Lens, a parameter-efficient adapter. Experimental results validated the effectiveness of both ECED and EEVIT. MACEE demonstrates strong generalizability, offering potential for expertise curation across medical domains in the future. ECED, as the first echocardiography expertise database, holds promise to inspire and support further advancements in echocardiography-specialized foundation models.

Acknowledgments. This work was supported by a research grant from the Joint Research Scheme (JRS) under the National Natural Science Foundation of China (NSFC) and the Research Grants Council (RGC) of Hong Kong (Project No. N_HKUST654/24), as well as a grant from the RGC of the Hong Kong Special Administrative Region, China (Project No. R6005-24). This study is validated by Dr. Hongwen Fei and Dr. Taoran Huang, two board-certified echocardiologists from Guangdong Cardiovascular Institute.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
2. Chen, J., Gui, C., Ouyang, R., Gao, A., Chen, S., Chen, G., Wang, X., Cai, Z., Ji, K., Wan, X., et al.: Towards injecting medical visual knowledge into multimodal llms at scale. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 7346–7370 (2024)
3. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
4. Christensen, M., Vukadinovic, M., Yuan, N., Ouyang, D.: Vision–language foundation model for echocardiogram interpretation. Nature Medicine pp. 1–8 (2024)
5. Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.E., Lomeli, M., Hosseini, L., Jégou, H.: The faiss library. arXiv preprint arXiv:2401.08281 (2024)
6. He, S., Nie, Y., Chen, Z., Cai, Z., Wang, H., Yang, S., Chen, H.: Meddr: Diagnosis-guided bootstrapping for large-scale medical vision-language learning. arXiv e-prints pp. arXiv–2404 (2024)
7. Hu, E.J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)

8. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
9. Kraljevic, Z., Searle, T., Shek, A., Roguski, L., Noor, K., Bean, D., Mascio, A., Zhu, L., Folarin, A.A., Roberts, A., et al.: Multi-domain clinical natural language processing with medcat: the medical concept annotation toolkit. Artificial intelligence in medicine **117**, 102083 (2021)
10. Lei, W., Ge, Y., Yi, K., Zhang, J., Gao, D., Sun, D., Ge, Y., Shan, Y., Shou, M.Z.: Vit-lens: Towards omni-modal representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26647–26657 (2024)
11. Lin, W., Zhao, Z., Zhang, X., Wu, C., Zhang, Y., Wang, Y., Xie, W.: Pmc-clip: Contrastive language-image pre-training using biomedical documents. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 525–536. Springer (2023)
12. Mitchell, C., Rahko, P.S., Blauwet, L.A., Canaday, B., Finstuen, J.A., Foster, M.C., Horton, K., Ogunyankin, K.O., Palma, R.A., Velazquez, E.J.: Guidelines for performing a comprehensive transthoracic echocardiographic examination in adults: recommendations from the american society of echocardiography. Journal of the American Society of Echocardiography **32**(1), 1–64 (2019)
13. Omerovic, S., Jain, A.: Echocardiogram (2020)
14. Rückert, J., Bloch, L., Brüngel, R., Idrissi-Yaghir, A., Schäfer, H., Schmidt, C.S., Koitka, S., Pelka, O., Abacha, A.B., G. Seco de Herrera, A., et al.: Rocov2: Radiology objects in context version 2, an updated multimodal image dataset. Scientific Data **11**(1), 688 (2024)
15. Vukadinovic, M., Tang, X., Yuan, N., Cheng, P., Li, D., Cheng, S., He, B., Ouyang, D.: Echoprime: A multi-video view-informed vision-language model for comprehensive echocardiography interpretation. arXiv preprint arXiv:2410.09704 (2024)
16. Wang, B., Xu, C., Zhao, X., Ouyang, L., Wu, F., Zhao, Z., Xu, R., Liu, K., Qu, Y., Shang, F., Zhang, B., Wei, L., Sui, Z., Li, W., Shi, B., Qiao, Y., Lin, D., He, C.: Mineru: An open-source solution for precise document content extraction (2024)
17. Wang, L., Wang, H., Yang, H., Mao, J., Yang, Z., Shen, J., Li, X.: Interpretable bilingual multimodal large language model for diverse biomedical tasks. In: The Thirteenth International Conference on Learning Representations (2025)
18. Wei, C., Milligan, M., Lam, M., Heidenreich, P.A., Sandhu, A.: Variation in cost of echocardiography within and across united states hospitals. Journal of the American Society of Echocardiography **36**(6), 569–577 (2023)
19. Wu, R., Su, N., Zhang, C., Ma, T., Zhou, T., Cui, Z., Tang, N., Mao, T., Zhou, Y., Fan, W., et al.: Mm-retinal v2: Transfer an elite knowledge spark into fundus vision-language pretraining. arXiv preprint arXiv:2501.15798 (2025)
20. Wu, R., Zhang, C., Zhang, J., Zhou, Y., Zhou, T., Fu, H.: Mm-retinal: Knowledge-enhanced foundational pretraining with fundus image-text expertise. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 722–732. Springer (2024)
21. Yang, H., Song, S., Qin, Y., Wang, L., Wang, H., Ding, X., Zhang, Q., Du, B., Li, X.: Multi-modal explainable medical ai assistant for trustworthy human-ai collaboration. arXiv preprint arXiv:2505.06898 (2025)
22. Yang, J., Ding, X., Zheng, Z., Xu, X., Li, X.: Graphecho: Graph-driven unsupervised domain adaptation for echocardiogram video segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11878–11887 (2023)

23. Yang, J., Huang, T., Ding, S., Xu, X., Zhao, Q., Jiang, Y., Guo, J., Pu, B., Zheng, J., Zhang, C., et al.: Ai-enabled accurate non-invasive assessment of pulmonary hypertension progression via multi-modal echocardiography. arXiv preprint arXiv:2505.07347 (2025)
24. Yang, J., Lin, Y., Pu, B., Guo, J., Xu, X., Li, X.: Cardiacnet: Learning to reconstruct abnormalities for cardiac disease assessment from echocardiogram videos. In: European Conference on Computer Vision. pp. 293–311. Springer (2024)
25. Yang, S., Du, J., Guo, J., Zhang, W., Liu, H., Li, H., Wang, N.: Vilref: An expert knowledge enabled vision-language retinal foundation model. arXiv preprint arXiv:2408.10894 (2024)
26. Zhang, X., Wu, C., Zhang, Y., Xie, W., Wang, Y.: Knowledge-enhanced visual-language pre-training on chest radiology images. Nature Communications **14**(1), 4542 (2023)
27. Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., Wang, Y., Xie, W.: Development of a large-scale medical visual question-answering dataset. Communications Medicine **4**(1), 277 (2024)
28. Zhou, X., Sun, L., He, D., Guan, W., Wang, R., Wang, L., Sun, X., Sun, K., Zhang, Y., Wang, Y., Xie, W.: A knowledge-enhanced pathology vision-language foundation model for cancer diagnosis. CoRR (2024)
29. Zhou, X., Zhang, X., Wu, C., Zhang, Y., Xie, W., Wang, Y.: Knowledge-enhanced visual-language pretraining for computational pathology. In: European Conference on Computer Vision. pp. 345–362. Springer (2024)
30. Zoghbi, W.A., Adams, D., Bonow, R.O., Enriquez-Sarano, M., Foster, E., Grayburn, P.A., Hahn, R.T., Han, Y., Hung, J., Lang, R.M., et al.: Recommendations for noninvasive evaluation of native valvular regurgitation: a report from the american society of echocardiography developed in collaboration with the society for cardiovascular magnetic resonance. Journal of the American Society of Echocardiography **30**(4), 303–371 (2017)