EchoVQA: Enabling Conversational Assistance for Point-of-Care Cardiac Ultrasound

Filippos Bellos[1], Yutong Li[1⋆], Jessie N. Dong[2⋆], Zaiyang Guo[2⋆], Emily Mackay[2], Yayuan Li[1], Yannis Avrithis[3], Alison Pouch[2], and Jason J. Corso[1]

[1] University of Michigan
{fbellos,jjcorso}@umich.edu
[2] University of Pennsylvania
[3] Independent Scientist

Abstract. Point-of-care transthoracic echocardiography enables cardiac assessment in virtually any clinical setting, yet its diagnostic utility remains constrained by the expertise required for image acquisition and interpretation. Visual question answering (VQA) offers a promising paradigm for bridging this expertise gap through interactive clinical assistance, but existing echocardiography VQA datasets are limited in scale, restricted to high-quality images, and only cover a few views. We introduce EchoVQA, the first large-scale VQA dataset for echocardiography, comprising 14,299 images and 74,819 question-answer pairs. The dataset integrates public sources (EchoNet-Dynamic, CAMUS) with our own point-of-care acquisitions from two handheld probes (Lumify, Clarius), spanning diverse views and including both high-quality and suboptimal images. Uniquely, EchoVQA includes acquisition guidance questions to help users optimize transducer positioning toward a diagnostic apical 4-chamber view for left ventricular ejection fraction estimation—a challenging task for novice operators in point-of-care settings. We further develop a parameter-efficient method based on multimodal learnable prompts achieving state-of-the-art performance on most benchmarks, including EchoVQA, with significantly fewer trainable parameters than existing state-of-the-art approaches. Data and code: https://huggingface.co/datasets/filbel/echoVQA.

Keywords: Point-of-Care Echocardiography · Visual Question Answering · Multimodal Large Language Models · VQA Benchmark.

1 Introduction

Transthoracic echocardiography (TTE) is the first-line imaging modality for assessing cardiac function due to its portability, lack of ionizing radiation, and ease of rapid acquisition [2]. Point-of-care (POC) TTE platforms—comprising handheld transducers connected to smartphones or tablets—have expanded cardiac imaging access to resource-limited settings [9]. However, interpreting echocardiographic images remains highly operator-dependent, requiring simultaneous identification of standard views, assessment of image quality, evaluation of chamber

⋆ Contributed equally.

visibility, and determination of suitability for downstream measurements such as left ventricular ejection fraction (LVEF). LVEF is a key metric of systolic function commonly assessed from the apical 4-chamber (A4C) view, yet optimizing transducer positioning for high-quality A4C acquisition is challenging for novice users [1, 17]. In POC settings, trained sonographers are often unavailable, and clinicians must acquire and interpret images with limited expertise and computational resources.

VQA has emerged as a powerful paradigm bridging vision and language processing to provide interactive assistance for medical image analysis [13]. Several medical benchmark datasets have been introduced: VQA-RAD [10] provides 3,515 QA pairs across radiology images, SLAKE [14] offers bilingual physician-verified annotations, PathVQA [5] contributes 32,799 pairs for pathology, and recent large-scale efforts such as MedTrinity-25M [26] provide millions of image-text pairs across diverse modalities. However, none specifically target echocardiography. To our knowledge, the only existing echocardiography VQA benchmark [23] provides just 622 QA pairs from high-quality hospital acquisitions, insufficient for training and lacking acquisition guidance for POC settings.

Enabling conversational assistance for medical imaging requires not only comprehensive datasets but also capable models. Multimodal large language models (MLLMs) are a promising solution. LLaVA-Med [12] adapts general-domain vision-language models to biomedicine through curriculum learning on figure-caption pairs followed by instruction-tuning. LLaVA-Tri [26] further improves performance through pretraining on MedTrinity-25M's multigranular annotations. However, these approaches require training 7–8B parameters, posing significant computational demands. Parameter-efficient fine-tuning (PEFT) methods aim to reduce this burden—PeFoMed [4] fine-tunes an MLLM with LoRA using only 33M trainable parameters—yet underperforms fully fine-tuned approaches.

In this work, we address these challenges with two contributions. First, we introduce EchoVQA, the largest VQA dataset for echocardiography, comprising 14,299 images and 74,819 question-answer pairs, featuring multi-turn conversations covering view classification, image quality assessment, chamber visibility, LVEF estimation feasibility, and acquisition guidance. Second, we propose a parameter-efficient method for adapting multimodal large language models to medical VQA, combining visual adaption prompts with text prompt learning [16]—previously unexplored in multimodal medical settings—and instance-level prompt optimization. Our method achieves state-of-the-art performance on most Med-VQA benchmarks with significantly fewer trainable parameters than LLaVA-Med (7B) and LLaVA-Tri (8B), while substantially outperforming PeFoMed.

2 EchoVQA Dataset

2.1 Data Collection and Pre-Processing.

We collect echocardiographic images from two complementary sources: public datasets and our own POC acquisitions.

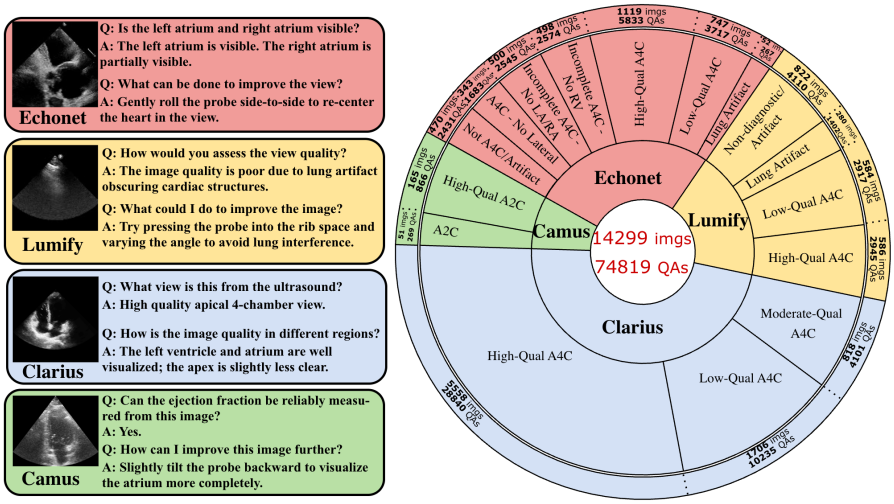

Fig. 1. EchoVQA dataset overview. **Left**: Example QA pairs from each data source demonstrating the different question types. **Right**: Distribution of images and QA pairs across data sources (inner ring) and quality categories (outer ring). The dataset comprises 14,299 images and 74,819 QA pairs spanning public (EchoNet-Dynamic, CAMUS) and our POC (Lumify, Clarius) acquisitions.

Public Data. Public datasets such as EchoNet-Dynamic [19] and CAMUS [11] are acquired by trained sonographers in hospital echo labs using cart-based systems. Our inspection reveals substantial variation within EchoNet-Dynamic, including non-standard views (e.g. A4C missing structures), lung artifacts, etc.) (Fig. 1). We explicitly examine and categorize these variations to enable targeted question-answer generation.

Point-of-Care Data. To complement public data with realistic POC acquisitions, typically performed with handheld probes by non-specialist clinicians, we collect additional scans using the Philips Lumify and Clarius PAL HD3, under institutional IRB approval. Using the Lumify, two operators performed scans on 3 subjects, and using the Clarius, scans were collected from 9 subjects (12 total; 7F/5M; age 23.5–40.6; 3 White, 9 Asian).These acquisitions capture the full spectrum of image quality a novice sonographer would encounter, stratified into five categories reflecting the progression toward a diagnostic A4C view for LVEF estimation: (1)*high-quality A4C*—all four chambers clearly visible, LVEF readily measurable; (2)*moderate-quality A4C*—partial chamber visibility with medium image quality, LVEF generally measurable; (3)*low-quality A4C*—recognizable A4C but compromised by suboptimal visualization, missing structures, off-axis orientation, or poor image settings, LVEF typically not feasible; (4)*lung artifact*—rib shadows or pleural interfaces obscuring cardiac structures; and (5) *non-diagnostic/artifacts*—inadequate for clinical interpretation due to severe arti-

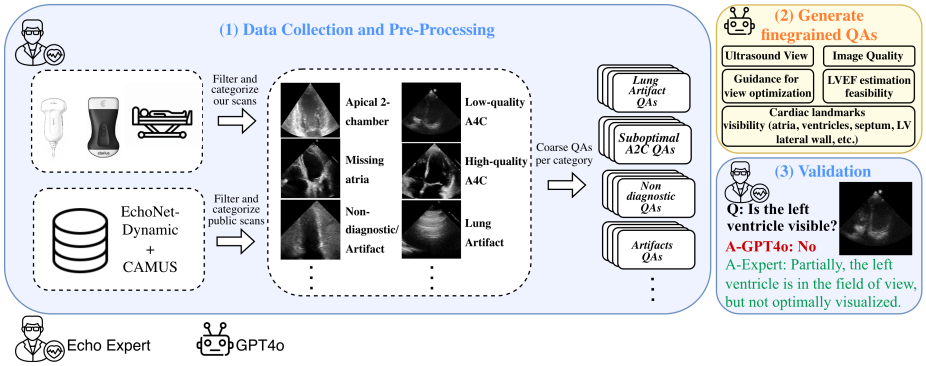

Fig. 2. EchoVQA construction pipeline. (1) Echocardiography experts categorize images from our POC acquisitions and public datasets into view and quality-based categories, then define coarse QA templates for each category. (2) GPT-4o generates diverse, image-specific QA pairs conditioned on category labels and expert templates, covering cardiac landmark visibility, LVEF estimation feasibility, ultrasound view classification, acquisition guidance, and image quality. (3) Experts review all generated QAs, correcting inaccurate or poorly grounded responses to ensure clinical accuracy.

facts or cardiac structures not in view. The specific nuances within each category are captured in the corresponding QA pairs. We adopt image-level rather than video-level QA because corrective guidance is frame-specific—differing across frames of a single clip—and composes with downstream video LVEF models; we target A4C as the canonical and hardest-to-acquire POC view for LVEF.

2.2 Question-Answer Generation.

We develop a human-in-the-loop annotation pipeline that ensures clinical accuracy while enabling efficient scaling through LLM assistance (Figure 2).

Category-Level Expert Templates. For each image category, echocardiography experts define a set of template question-answer pairs that establish the clinical scope and ground truth for that category. These templates encode expert knowledge about quality assessment, expected chamber visibility, LVEF estimation feasibility, and acquisition guidance—ranging from correcting suboptimal views toward a diagnostic A4C (e.g., "Adjust the probe position laterally and angle it toward the patient's left shoulder") to further optimizing already diagnostic images (e.g., "Rotate the probe slightly clockwise to better visualize the septum and right atrium"). These expert-authored templates serve as conditioning context for scalable QA generation.

Fine-Grained QA Generation. Using GPT-4o, we expand each category's expert-defined templates into diverse, image-specific question-answer pairs, conditioning on the category label, the templates, and the image. Categories are assigned by experts, not the model. Critically, while templates provide clinical grounding, the model is prompted to analyze each image directly, as substantial

variability exists even within the same category. For instance, images categorized as low-quality A4C may exhibit, among other issues, low contrast obscuring chamber boundaries, partial left ventricular cutoff compromising LVEF estimation while other chambers remain visible, or foreshortening that distorts the true apex position. The model generates questions spanning five types for each ultrasound image: view classification, chamber visibility assessment, image quality evaluation, LVEF estimation feasibility, and acquisition guidance for view optimization. This approach enables efficient scaling—generating thousands of QA pairs per category—while ensuring all generated content remains anchored to both expert-validated clinical knowledge and image-specific observations.

2.3 Expert Validation.

Although MLLMs have shown promise in medical imaging, they remain imperfect at specialized medical tasks and may introduce systematic biases in generated content. It is therefore imperative that all generated QA pairs undergo expert review. All 74,819 QA pairs were reviewed by four echocardiography experts; we prioritized full-coverage correctness review over inter-rater agreement on a subset. For each pair, reviewers verified factual correctness, ensured answers were grounded in the specific image rather than reflecting generic category-level knowledge, and corrected or rewrote responses that failed to meet these criteria. They further refined answer phrasing for clinical precision and rejected questions that were ambiguous or inappropriate for a given image. This validation step ensures that despite LLM-assisted generation, the final dataset meets the quality standards expected for clinical use.

3 Parameter-Efficient Medical VQA

Current medical VQA methods operate at a single level of adaptation—modifying either full model weights or low-rank subspaces uniformly. We instead propose a parameter-efficient multi-level prompt framework that injects task-relevant information at three complementary levels: (1) visual adaption prompts [27] that inject visual information into the language model, (2) dataset-level text prompts that encode shared domain knowledge via prompt learning [16]—previously unexplored in multimodal medical settings, and (3) instance-level prompts optimized at inference to capture image-specific context.

3.1 Model Architecture

The proposed method and model architecture are illustrated in Figure 3.

Vision Encoder. The input echocardiography image X^v is encoded by a vision encoder f^v to obtain visual features $Z^v := f^v(X^v) \in \mathbb{R}^{D \times N_v}$, where D is the embedding dimension and N_v is the number of visual tokens. These features are projected to the LLM embedding space via a learnable projection $W_p \in \mathbb{R}^{D_L \times D}$, yielding $\tilde{Z}^v := W_p Z^v \in \mathbb{R}^{D_L \times N_v}$.

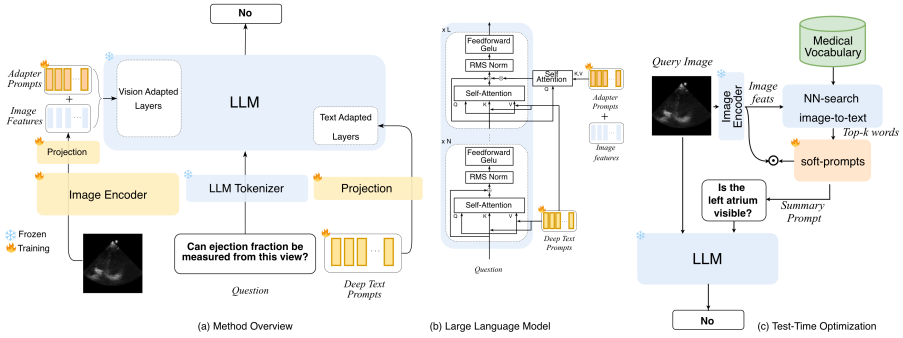

Fig. 3. Proposed method. **(a)** Training pipeline with visual adaption prompts injected into top LLM layers. **(b)** Dataset-level text prompts added to keys and values in the layer's self-attention. **(c)** At inference, instance-level prompts are retrieved from a medical vocabulary, optimized, and injected as a summary prompt.

Text Embedding. The input question X^q is tokenized into a sequence of S tokens and mapped by the frozen LLM embedding layer f^t to obtain text embeddings $Z^t := f^t(X^q) = (z_1^t, \ldots, z_S^t) \in \mathbb{R}^{D_L \times S}$.

Visual Adaption Prompts. Following [27], we inject visual information into the LLM via learnable adaption prompts $P^a \in \mathbb{R}^{D_L \times N_a}$. The projected visual features are pooled into a global token and added to P^a, then prepended to input sequences at the top L' transformer layers. We use zero-initialized gating with learnable scalars $\alpha^{(l)}$ for stable training.

Dataset-Level Text Prompts. Beyond visual adaption prompts, we introduce learnable text prompts that encode domain-specific knowledge shared across the dataset. These prompts are prepended to the keys and values in the self-attention mechanism. Given an input sequence $H \in \mathbb{R}^{D_L \times K}$, the query, key, and value are computed as: $Q := W^Q H, \quad K := [P^K; W^K H], \quad V := [P^V; W^V H]$, where $W^Q, W^K, W^V \in \mathbb{R}^{D_L \times D_L}$ are the projection matrices, and $P^K, P^V \in \mathbb{R}^{D_L \times N_p}$ are learnable text prompt parameters with N_p prompt tokens. This formulation allows the model to attend to learned domain-specific context without modifying the output sequence length.

3.2 Training

We adopt a two-stage training strategy following prior work in medical multimodal learning [4, 12, 26]. In **Stage 1**, we train on medical image-caption datasets to align visual representations with the language model's embedding space. In **Stage 2**, we fine-tune on downstream medical VQA datasets. The model is trained with the standard language modeling objective:

$$\mathcal{L} = -\sum_{i=1}^{|X^a|} \log p(x_i^a | X^v, X^q, x_{<i}^a; \Theta), \tag{1}$$

where Θ denotes all trainable parameters and $x^a_{<i}$ represents answer tokens preceding position i. During both stages, we update the vision encoder f^v, visual projection W_p, adaption prompts P^a, gating scalars $\{\alpha^{(l)}\}$, and text prompt key-value pairs $\mathcal{P}^t$, while keeping the LLM frozen.

3.3 Inference: Instance-Level Prompt Optimization

We introduce instance-level test time prompt optimization to inject image-specific context beyond the dataset-level prompts learned during training.

We construct a **medical vocabulary** $\mathcal{V} = \{v_1, v_2, \ldots, v_M\}$ containing M domain-specific terms from the train set of each downstream dataset. Each term v_m is encoded using BiomedCLIP's text encoder to obtain vocabulary embeddings $E^{\mathcal{V}} \in \mathbb{R}^{D \times M}$. Given a query image X^v, we retrieve the top-k most similar vocabulary terms via cosine similarity: $\mathcal{R} = \text{top-}k\left(\frac{Z^v \cdot E^{\mathcal{V}}}{\|Z^v\| \|E^{\mathcal{V}}\|}\right)$, where $\mathcal{R} = \{r_1, \ldots, r_k\}$ denotes the retrieved terms.

The retrieved terms $\mathcal{R}$ are tokenized and mapped to the LLM embedding space via the frozen embedding layer to obtain $P^{init} := f^t(\mathcal{R}) \in \mathbb{R}^{D_L \times N_r}$, where N_r is the total number of tokens. These embeddings initialize soft prompts P^{soft}, which are **optimized for T iterations** to better align with the visual features: $P^{soft*} = \arg\min_{P^{soft}} \mathcal{L}_{align}(P^{soft}, Z^v)$, where $\mathcal{L}_{align}$ is a cosine similarity loss. After optimization, we select the first retrieved term's embedding as a compact summary prompt capturing the dominant visual semantics. This single summary prompt is concatenated with the question embeddings $\hat{Z}^t = [P^{summary}; Z^t]$, and the final answer is generated by the LLM conditioned on visual features $\tilde{Z}^v$ and the augmented text embeddings $\hat{Z}^t$. The per-query overhead of test-time optimization consists of a single vocabulary retrieval and T forward passes through the frozen LLM to optimize the soft prompt, incurred once per image and independent of answer length.

4 Experiments

Implementation Details. We use BiomedCLIP [28] as our vision encoder and LLaMA-2-7B [24] as the language model. The LLM remains frozen during training while the vision encoder is fine-tuned. We set the number of visual adaption prompts $N_a = 10$ and text prompt tokens $N_p = 10$.

In Stage 1, we train on medical image-caption datasets (ROCO [20], ImageCLEF [7], MEDICAT [22], and MIMIC-CXR [8]). In Stage 2, we fine-tune on each downstream VQA dataset following [12, 26, 4]. We partition EchoVQA into 70/10/20 train/validation/test splits, stratified by data source and view-quality category to preserve the category and quality distribution. We report all results on the held-out test split. For test-time prompt optimization, we retrieve $k = 10$ terms from the medical vocabulary and optimize for $T = 100$ iterations. All experiments are conducted on 2 NVIDIA H100 GPUs.

Evaluation. We evaluate on three established medical VQA benchmarks: VQA-RAD [10], SLAKE [14], and PathVQA [5]. We additionally evaluate on our

Table 1. Performance comparison on medical VQA benchmarks. We report recall for open-ended questions and exact match accuracy for closed-ended questions. All methods except GPT-4V have been reproduced by us, using their respective codebases.

Method	Params	VQA-RAD		SLAKE		PathVQA		EchoVQA	
		Open	Closed	Open	Closed	Open	Closed	Open	Closed
GPT-4V	–	39.5	78.9	33.6	43.6	–	–	–	–
LLaVA	7B	63.7	83.4	83.7	84.6	37.1	90.8	56.0	70.6
PeFoMed (LoRA)	33M	58.5	83.4	81.3	82.7	29.5	88.4	54.4	86.9
LLaVA-Med	7B	64.7	84.2	83.0	85.3	**38.5**	91.2	55.5	82.2
LLaVA-Tri	8B	65.1	84.9	82.3	83.2	37.8	**91.8**	55.2	79.7
Ours	104M	**65.8**	**86.8**	**87.3**	**89.7**	37.2	91.3	**60.2**	**90.4**

proposed EchoVQA benchmark. Following [12, 26], we report recall for open-ended questions and exact match accuracy for closed-ended questions.

We compare against several state-of-the-art **baselines**: GPT-4V [18] as a zero-shot baseline; LLaVA [15] as a general-domain MLLM; LoRA finetuning (PeFoMed) [6, 4], LLaVA-Med [12] and LLaVA-Tri [26]. Echo-specific models (EchoCLIP [3], EchoPrime [25]) are contrastive/retrieval encoders rather than generative VQA models, hence not comparable under our generative exact-match protocol, while recent generalist medical MLLMs (e.g., MedGemma [21]) fold benchmark splits into multi-stage tuning under looser LLM-as-judge metrics, precluding apples-to-apples comparison.

Results. Table 1 presents results on established medical VQA benchmarks. Our method achieves state-of-the-art performance on VQA-RAD, SLAKE and EchoVQA, outperforming all baselines. On PathVQA we are competitive but not state-of-the-art. The likely cause is a domain gap: our radiology-focused Stage 1 pretraining transfers less well to PathVQA's diverse pathology images, and unlike fully fine-tuned baselines our frozen LLM relies more on the alignment learned during pretraining.

Table 2. Ablation study evaluating the contribution of visual adaption prompts (AP), dataset-level text prompts (TP), and test-time optimization (TTO).

Method	VQA-RAD		SLAKE		PathVQA		EchoVQA	
	Open	Closed	Open	Closed	Open	Closed	Open	Closed
AP	65.6	80.6	87.1	88.1	36.9	90.3	60.0	88.7
AP + TP	65.6	85.7	**87.3**	89.4	**37.2**	91.2	60.0	**90.4**
AP + TP + TTO (Full)	**65.8**	**86.8**	**87.3**	**89.7**	**37.2**	**91.3**	**60.2**	**90.4**

Table 2 isolates the contribution of each component. Visual adaption prompts alone provide a strong baseline. Adding dataset-level text prompts yields con-

sistent gains on closed-ended questions across all benchmarks, with the largest improvement on VQA-RAD (+5.1 closed). Test-time optimization provides further incremental gains.

5 Conclusion and Future Work

We introduced EchoVQA, the first large-scale VQA dataset for echocardiography, spanning diverse image-quality levels and acquisition guidance for POC settings, together with a parameter-efficient method that achieves state-of-the-art performance on EchoVQA and most standard medical VQA benchmarks. Future work includes augmenting the POC cohort with additional subjects spanning broader demographics in order to enable better generalization, a novice-operator user study assessing acquisition-guidance efficacy, inter-rater agreement analysis, full FLOPs/latency profiling of test-time optimization, and extension beyond A4C to additional standard views.

Acknowledgments. This research was funded, in part, by the U.S. Government, under Agreement No. 1AY2AX000062. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

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

References

1. Bick, J.S., Demaria, S., Kennedy, J.D., Schwartz, A.D., Weiner, M.M., Levine, A.I., Shi, Y., Schildcrout, J.S., Wagner, C.E.: Comparison of expert and novice performance of a simulated transesophageal echocardiography examination. Simulation in Healthcare: Journal of the Society for Simulation in Healthcare **8**(5), 329–334 (Oct 2013)
2. Cheitlin, M.D., et al.: ACC/AHA Guidelines for the Clinical Application of Echocardiography. Circulation **95**(6), 1686–1744 (Mar 1997), publisher: American Heart Association
3. Christensen, M., Vukadinovic, M., Yuan, N., Ouyang, D.: Vision–language foundation model for echocardiogram interpretation. Nature Medicine **30**(5), 1481–1488 (2024)
4. He, J., Li, P., Liu, G., He, G., Chen, Z., Zhong, S.: Pefomed: Parameter efficient fine-tuning of multimodal large language models for medical imaging. arXiv preprint arXiv:2401.02797 (2024)
5. He, X., Cai, Z., Wei, W., Zhang, Y., Mou, L., Xing, E., Xie, P.: Towards visual question answering on pathology images. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). pp. 708–718. Association for Computational Linguistics, Online (Aug 2021)

6. Hu, E.J., et al.: Lora: Low-rank adaptation of large language models. In: International Conference on Learning Representations (ICLR) (2022)
7. Ionescu, B., et al.: Overview of the imageclef 2022: Multimedia retrieval in medical, social media and nature applications. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Lecture Notes in Computer Science, vol. 13390, pp. 541–564. Springer, Cham (2022)
8. Johnson, A.E.W., et al.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data **6**, 317 (2019)
9. Kornelsen, J., Ho, H., Robinson, V., Frenkel, O.: Rural family physician use of point-of-care ultrasonography: experiences of primary care providers in British Columbia, Canada. BMC Primary Care **24**, 183 (Sep 2023)
10. Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Scientific Data **5**, 180251 (2018)
11. Leclerc, S., et al.: Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. IEEE Transactions on Medical Imaging **38**(9), 2198–2210 (2019)
12. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In: Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track (2023), spotlight
13. Lin, Z., Zhang, D., Tao, Q., Shi, D., Haffari, G., Wu, Q., He, M., Ge, Z.: Medical visual question answering: A survey. Artificial Intelligence in Medicine **143**, 102611 (2023)
14. Liu, B., Zhan, L.M., Xu, L., Ma, L., Yang, Y., Wu, X.M.: Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 1650–1654. IEEE (2021)
15. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
16. Liu, X., Ji, K., Fu, Y., Tam, W.L., Du, Z., Yang, Z., Tang, J.: P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 61–68. Association for Computational Linguistics (2022)
17. Mitchell, C., et al.: Guidelines for Performing a Comprehensive Transthoracic Echocardiographic Examination in Adults: Recommendations from the American Society of Echocardiography. Journal of the American Society of Echocardiography: Official Publication of the American Society of Echocardiography **32**(1), 1–64 (Jan 2019)
18. OpenAI, et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
19. Ouyang, D., et al.: Video-based AI for beat-to-beat assessment of cardiac function. Nature **580**(7802), 252–256 (Apr 2020), publisher: Nature Publishing Group
20. Pelka, O., Koitka, S., Rückert, J., Nensa, F., Friedrich, C.M.: Radiology objects in context (roco): A multimodal image dataset. In: Multimodal Learning for Clinical Decision Support and Clinical Image-Based Procedures (LABELS / CVII-STENT @ MICCAI). Lecture Notes in Computer Science, vol. 11043, pp. 180–189. Springer, Cham (2018)
21. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., et al.: MedGemma technical report (2026)

22. Subramanian, S., et al.: Medicat: A dataset of medical images, captions, and textual references. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 2112–2120. Association for Computational Linguistics, Online (2020)
23. Thapa, R., Li, A., Wu, Q., He, B., Sahashi, Y., Binder-Rodriguez, C., Zhang, A., Ouyang, D., Zou, J.: MIMIC-IV-ECHO-Ext-MIMICEchoQA: A Benchmark Dataset for Echocardiogram-Based Visual Question Answering. PhysioNet (Oct 2025), version 1.0.0
24. Touvron, H., et al.: Llama 2: Open Foundation and Fine-Tuned Chat Models. ArXiv (Jul 2023)
25. Vukadinovic, M., Chiu, I.M., Tang, X., Yuan, N., Chen, T.Y., Cheng, P., Li, D., Cheng, S., He, B., Ouyang, D.: Comprehensive echocardiogram evaluation with view-primed vision-language AI. Nature **650**(8103), 970–977 (2026)
26. Xie, Y., Zhou, C., Gao, L., Wu, J., Li, X., Zhou, H.Y., Liu, S., Xing, L., Zou, J., Xie, C., Zhou, Y.: Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine. In: International Conference on Learning Representations (ICLR) (2025)
27. Zhang, R., et al.: LLaMA-Adapter: Efficient Fine-tuning of Language Models with Zero-init Attention (2023), publisher: arXiv Version Number: 3
28. Zhang, S., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)