V2T-CoT: From Vision to Text Chain-of-Thought for Medical Reasoning and Diagnosis

Yuan Wang[1,2,6⋆], Jiaxiang Liu[1*], Shujian Gao[3], Bin Feng[4], Zhihang Tang[5], Xiaotang Gai[1], Jian Wu[1,2], and Zuozhu Liu[1,2✉]

[1] Zhejiang University-University of Illinois Urbana-Champaign Institute, Zhejiang University, Zhejiang, China
[2] Zhejiang Key Laboratory of Medical Imaging Artificial Intelligence, Zhejiang, China
[3] Academy for Engineering and Technology, Fudan University, Shanghai, China
[4] Department of Oral and Maxillofacial Radiology, Stomatology Hospital, School of Stomatology, Zhejiang University, Zhejiang, China
[5] Intelligent Computing Infrastructure Innovation Center, Zhejiang Lab, Zhejiang, China
[6] ChohoTech Inc., Zhejiang, China
{yuan2.24,zuozhuliu}@intl.zju.edu.cn
https://github.com/Venn2336/V2T_CoT

Abstract. Recent advances in multimodal techniques have led to significant progress in Medical Visual Question Answering (Med-VQA). However, most existing models focus on global image features rather than localizing disease-specific regions crucial for diagnosis. Additionally, current research tends to emphasize answer accuracy at the expense of the reasoning pathway, yet both are crucial for clinical decision-making. To address these challenges, we propose From Vision to Text Chain-of-Thought (**V2T-CoT**), a novel approach that automates the localization of preference areas within biomedical images and incorporates this localization into region-level pixel attention as knowledge for Vision CoT. By fine-tuning the vision language model on constructed **R-Med 39K** dataset, V2T-CoT provides definitive medical reasoning paths. V2T-CoT integrates visual grounding with textual rationale generation to establish precise and explainable diagnostic results. Experimental results across four Med-VQA benchmarks demonstrate state-of-the-art performance, achieving substantial improvements in both performance and interpretability.

Keywords: Med-VQA · Vision Language Model · Chain of Thought.

1 Introduction

Medical Visual Question Answering (Med-VQA) has emerged as a critical domain in healthcare AI, leveraging multimodal deep learning to answer clinical questions related to images posed in natural language [5,8,10,12]. Med-VQA

⋆ Equal contribution. ✉Corresponding author.

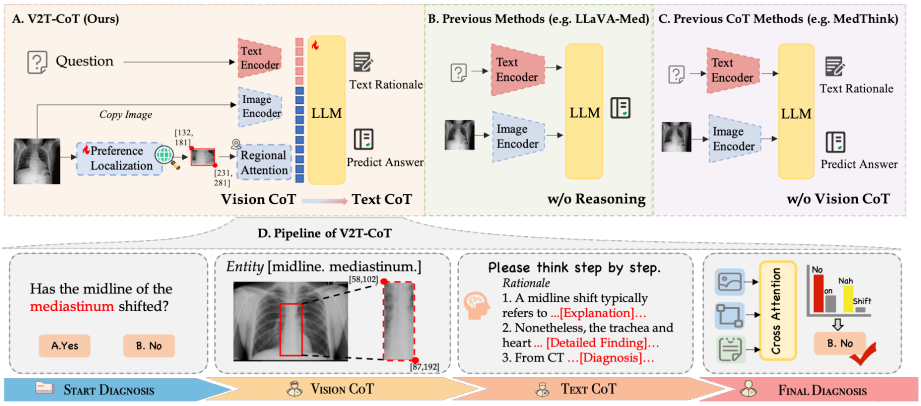


Fig. 1: The comparison between V2T-CoT and existing Med-VQA methods. **A** employs a combined vision and text encoding strategy with regional attention for medical diagnosis. In contrast, previous methods (**B** & **C**) either lack reasoning or utilize it in a text-only context. **D** demonstrates the pipeline of V2T-CoT.

holds promise for automated, personalized health consultations, addressing the growing demand for accessible and timely healthcare [20]. By integrating textual data (e.g., clinical notes, patient histories) with medical images (e.g., X-rays, MRIs), these models provide a holistic understanding of patient conditions, improving decision-making and care delivery.

Recent advancements in Med-VQA have adopted deep learning frameworks such as DoctorGLM [27],Huatuo-o1 [4] and MedFound [22] to enhance the medical reasoning process. However, three core challenges persist [29,21]: **First**, existing feature fusion mechanisms (e.g., concatenation/weighted operations) suffer from representational capacity limitations. These coarse-grained fusion strategies fail to establish precise mappings between visual key regions (e.g., lesion areas in CT/MRI) and textual semantics, resulting in ineffective capture of diagnosis-relevant fine-grained cross-modal correlations. **Second**, high-accuracy models lack explainable reasoning pathways. Taking LLaVA-Med [16] (Fig. 1B) as an example, despite achieving high accuracy, its black-box decision mechanism cannot generate clinical reasoning chains (e.g., progressive analysis for malignancy determination based on tumor morphological features), severely hindering trustworthiness verification in diagnostic scenarios [21]. **Third**, multimodal CoT methods face both data and algorithm constraints. Frameworks like MedThink [7] (Fig. 1C), while incorporating clinical reasoning chains, are limited by: annotation data scarcity due to manual labeling costs; absence of specialized instruction-tuning datasets for training reasoning paths [25,21]. These limitations collectively lead to insufficient generalization capabilities in multi-perspective complex medical analysis.

In this paper, we propose a novel Med-VQA method: From Vision to Text CoT (**V2T-CoT**), which addresses the aforementioned challenges by effectively

integrating textual and visual reasoning (Fig. 1A). This method combines visual clues with textual information, comprising two core components: (1) a preference localization mechanism that identifies key regions in descriptive text relevant to the image; (2) a regional pixel-level attention mechanism that focuses on specific areas of medical images crucial for final diagnosis. To further enable explainable reasoning, we construct an instruction-tuning dataset **R-Med 39K** containing multi-granularity reasoning paths. Ultimately, our vision language model (VLM) generates coherent and interpretable diagnostic rationales, from which medical conclusions are derived. Experiments demonstrate that this approach achieves SOTA performance under comparable parameter sizes, significantly outperforming existing models. The main contributions are summarized as follows:

- We propose a medical visual to text chain-of-thought reasoning framework, **V2T-CoT**, which provides disease-related visual cues and a clear reasoning path, facilitating medical diagnosis.
- To locate disease-related visual cues, V2T-CoT introduces an automated method (Vision CoT) for identifying and aligning relevant image regions.
- We constructed **R-Med 39K** (Instruction-Tuning dataset) with reasoning paths from four Med-VQA datasets for Text CoT, integrating them into VLM training for reliable rationale validated by both LLMs and experts.
- Through extensive experiments, we validated the superior performance of V2T-CoT on four Med-VQA datasets compared to existing methods, demonstrating the effectiveness of Vision CoT for visual localization and the rationale in Text CoT for reasoning pathways, meeting the clarity and interpretability requirements of both humans and LLMs.

2 Method

2.1 Overview

In this work, we propose a novel multimodal approach for Med-VQA that integrates visual and textual reasoning chains. As shown in Fig. 1D, V2T-CoT first leverages a vision encoder to extract spatial-semantic features from critical anatomical regions, explicitly capturing structural deviations such as midline shifts (Vision CoT). The reasoning module then synthesizes anatomical region proposals with multimodal evidence to construct diagnostic rationales (Text CoT), ultimately delivering clinically interpretable diagnoses with both accuracy and clinical diagnostic transparency.

2.2 Vision CoT

Formulation of Phrase Grounding. We reformulate medical object detection as phrase grounding within the Vision CoT framework, aligning each image region with text prompt phrases. As shown in Fig. 2, given an image and text prompt, visual encoder Enc_I extracts region features $V \in \mathbb{R}^{N \times d}$, while language encoder Enc_L encodes textual tokens $T \in \mathbb{R}^{M \times d}$. The alignment score matrix $S_{\text{ground}} \in \mathbb{R}^{N \times M}$ is computed via: $S_{\text{ground}} = VT^{\top}$, where V and T denote region-level visual features and phrase-level text embeddings respectively.

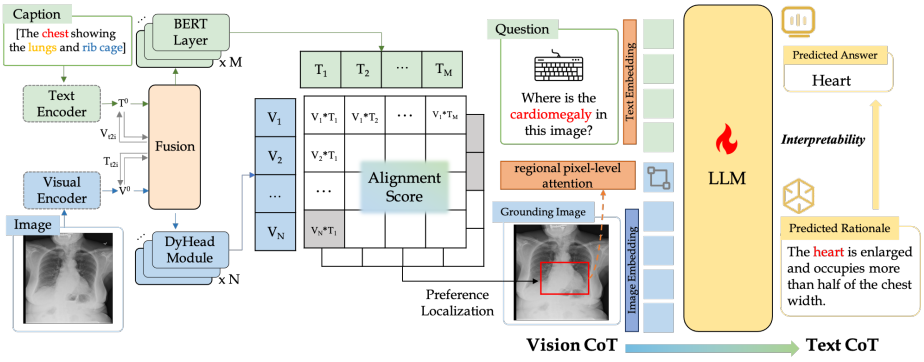


Fig. 2: Implementation of V2T-CoT for medical diagnosis. The model leverages dynamic fusion of image and text encodings, alignment scoring, and regional attention mechanisms to generate answers and rationale.

Fusion of Visual and Language Features. To enhance the interaction between visual and textual modalities, we introduce a fusion mechanism in the last layers of the encoders. This mechanism involves cross-modality multi-head attention (X-MHA) modules that facilitate the exchange of information between the image and text encoders. The fused encoder is defined as:

$$V^i_{\text{t2i}}, T^i_{\text{i2t}} = \text{X-MHA}(V^i, T^i), \quad i \in \{0, 1, .., L-1\}, \tag{1}$$

$$V^{i+1} = \text{DyHeadModule}(V^i + V^i_{\text{t2i}}), \quad V = V^L, \tag{2}$$

$$T^{i+1} = \text{BERTLayer}(T^i + T^i_{\text{i2t}}), \quad T = T^L, \tag{3}$$

where L is the number of DyHeadModules, V^i_{t2i} and T^i_{i2t} are the context vectors generated by the X-MHA module.

Regional Pixel-Level Attention. Following Preference Localization, we refine the model's focus through regional pixel-level attention, enhancing its ability to target specific areas crucial for diagnosing conditions such as cardiomegaly.

Regional pixel-level attention is implemented by assigning higher weights to key pixels using convolutional layers and attention modules. To balance regional and global attention, we introduce a weight factor α, where $\alpha \in [0, 1]$. The combined attention A_{combined} is computed as:

$$A_{\text{combined}} = \text{SoftMax}\left(\alpha \cdot Q_{\text{regional}} K^\top_{\text{regional}} + (1-\alpha) \cdot Q_{\text{global}} K^\top_{\text{global}}\right) / \sqrt{d}, \tag{4}$$

where $Q_{\text{regional}} \in \mathbb{R}^{h \times w \times d}$ and $K_{\text{regional}} \in \mathbb{R}^{h \times w \times d}$ are the query and key features from the localized region, while $Q_{\text{global}} \in \mathbb{R}^{H \times W \times d}$ and $K_{\text{global}} \in \mathbb{R}^{H \times W \times d}$ are the query and key features from the global image. The weight α controls the balance between the local and global attention, allowing the model to focus on fine-grained details while considering broader contextual information.

2.3 Text CoT

R-Med 39K Dataset Construction. The text rationale construction process adopts a two-stage generation and verification pipeline. Initially, given the input tuple $< Question(Q), Image(I) >$, a vision language model generates the preliminary answer(A). Subsequently, leveraging the triplet $< Q, I, A >$, we employ a reasoning prompting strategy to produce a structured diagnostic rationale(R). To ensure clinical validity, the generated R undergoes verification where a separate language model evaluates its logical consistency with A and I. This process yields the final quadruple $< Q, I, A, R >$ for instruction tuning. As demonstrated in Figure 4, quality control combines automated metrics with expert validation, with implementation details provided in Section 3.3.

Implementation of the Text CoT. Given the visual features $V \in \mathbb{R}^{N\times d}$ extracted from the Vision CoT and the textual features $T \in \mathbb{R}^{M\times d}$ encoded by the language model, the Text CoT module computes the cross-modal alignment scores as: $S_{\text{text}} = \text{SoftMax}\left(\frac{Q_{\text{text}}K_{\text{text}}^{\top}}{\sqrt{d}}\right)$, where Q_{text} and K_{text} are the query and key projections of the textual and visual features, respectively.

To further enhance the interaction between modalities, we employ a multi-head cross-attention mechanism in the VLM, where the textual features are iteratively refined through layers of self-attention:

$$T^{i+1} = \text{MultiHeadAttention}(T^i, V_{regional}, V_{global}), \tag{5}$$

where T^i represents the textual features at layer i. This iterative refinement ensures that the textual reasoning aligns with the visual clues.

3 Experiments

3.1 Experiments Setting

Datasets. Four benchmarks are used as the training and testing datasets for V2T-CoT. Among them, VQA-RAD [15], SLAKE [18], and VQA-2019 [3] are based on radiology images (e.g., CT, MRI), while PathVQA [11] is built on pathological images (e.g., tissue slides). Following Section 2.3, we generated rationales (thinking pathway) via LLMs (GPT-4 and Gemini), and construct them into instruction-tuning dataset R-Med containing 39K quadruple$< Q, I, A, R >$ format. In accordance with the original data partition, the questions are categorized into closed-ended (single-answer) and open-ended (free-form) types [21].

Implementation Details. Vision CoT initializes with the GLIP [17] detector, fine-tuned on SLAKE to adapt it to the medical domain. It is then deployed in an automated pipeline to generate region proposals for other datasets without manual annotation. For VLM, we employ CLIP-ViT-L/14@336px as the vision encoder to extract features from input images, distinguishing between global

attention and regional pixel-level attention. For the language model, StableLM (1.6B) and Phi2 (2.7B) are utilized [2,9]. Additionally, a 2-layer Multi-Layer Perceptron (MLP) is implemented as the projector to facilitate the integration of visual and textual features.

In the pretraining phase, the 2-layer multimodal projector is trained on a large-scale dataset of medical image-caption pairs [16], establishing a foundational understanding of visual data [19]. In the instruction fine-tuning phase, the CLIP encoder is frozen, and the remaining model components, including the visual encoder and language model, are fine-tuned on our R-Med dataset consisting of up to 39K quadruple. Both pretraining and fine-tuning are performed on 4 NVIDIA GeForce RTX 3090 GPUs.

Table 1: Comparison of model performances across different datasets on closed-end question: The **bolded** and underlined values correspond to the best and 2nd best performance, respectively. All results are in %.

Method	w/ MLLM	Act.	VQA-RAD	SLAKE	VQA-2019	PathVQA
Representative & SoTA methods with numbers reported in the literature						
VL Encoder-Decoder [1]		-	82.47	-	-	85.61
Q2ATransformer [23]		-	81.20	-	-	88.5
MMBERT [14]		-	77.90	-	78.10	-
PubMedCLIP [6]		-	80.00	82.50	-	-
BiomedGPT-M [28]		-	65.07	86.80	-	85.70
Prefix T. Medical LM [26]	✓	1.5B	-	82.01	-	87.00
Gemini Pro [24]	✓	-	60.29	72.60	60.22	70.30
Supervised finetuning results						
MedThink [7]		223M	83.50	86.30	-	87.00
LLaVA [19]	✓	7B	65.07	63.22	-	63.20
LLaVA-Med(LLama) [16]	✓	7B	<u>84.19</u>	85.34	-	91.21
LLaVA-Med(Vicuna) [16]	✓	7B	81.98	83.17	-	**91.65**
LLaVA-Med(Phi2) [16]	✓	2.7B	81.35	83.29	-	90.17
Med-MoE [13]	✓	2.0B	80.07	83.41	76.56	91.30
V2T-CoT (Stablelm)	✓	1.6B	80.08	<u>86.48</u>	<u>79.69</u>	90.42
V2T-CoT(Phi2)	✓	2.7B	**84.86**	**87.61**	**80.10**	<u>91.42</u>

3.2 Main Results

In the experiment, we compared mainstream methods on four Med-VQA datasets, using accuracy as the metric for closed-end questions. As shown in Table 1, V2T-CoT (Phi2) achieved the highest accuracy on VQA-RAD and SLAKE, demonstrating superior performance. A significant improvement over the baseline results can be observed, with an increase of 2.39% on VQA-RAD and 5.11%

on SLAKE, respectively, outperforming LLaVA-Med (Vicuna) [16]. Additionally, it attained competitive results on VQA-2019 and PathVQA, surpassing the compared method by 2.00% over MMBERT [14] and 1.25% over LLaVA-Med (Phi2) [16], respectively. For open-end questions, Fig. 3A provides a comparison of the performance of V2T-CoT and MedThink [7] across three datasets, using open-ended evaluation metrics such as BLEU and Rouge-L. Results show that V2T-CoT consistently outperforms MedThink across all metrics in three datasets. Specifically, V2T-CoT highlights superior performance in BLEU-1 and Rough-L, highlighting its enhanced understanding and generation capabilities.

Table 2: Ablation Study on the Impact of Vision and Text CoT Integration: Performance Comparison on Closed-end and Open-end Questions. The **bolded** and underlined values correspond to the best and 2nd best performing results.

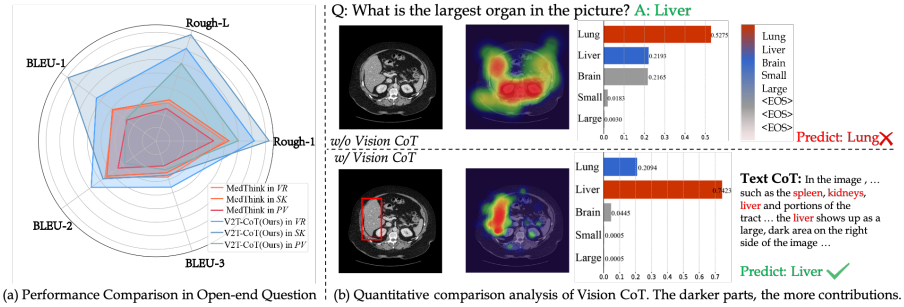
Vision CoT	Text CoT	Closed-end Question (Accuracy)			Open-end Question (Recall)		
		VQA-RAD	SLAKE	PathVQA	VQA-RAD	SLAKE	PathVQA
		<u>83.82</u>	84.38	<u>91.33</u>	56.45	82.73	29.75
	✓	82.34	86.48	90.83	58.72	<u>83.24</u>	<u>31.57</u>
✓		83.27	<u>87.04</u>	<u>91.33</u>	<u>59.53</u>	80.24	29.70
✓	✓	**84.86**	**87.61**	**91.42**	**60.40**	**84.08**	**31.62**

3.3 Ablation Analysis

Table 2 demonstrates the necessity of integrating visual and textual reasoning chains in V2T-CoT. V2T-CoT outperforms all datasets, showcasing the complementary roles of its components. Vision CoT improves spatial reasoning, reducing SLAKE's closed-end question error by 2.66%. Text CoT is crucial for logical coherence in open-ended tasks, with its absence causing a 2.27% drop in VQA-RAD open-ended recall.

Case Study Analysis. Fig. 3B demonstrates Vision CoT's capability in lesion-centric attention guidance, where heatmap precisely localize anatomical structures (e.g., liver identification as the largest organ), thereby improving reasoning accuracy and diagnostic interpretability through targeted visual grounding.

Vision CoT Regional Detection Analysis. To evaluate the impact of Vision CoT's localization capability on Med-VQA task, we compare various object detection algorithms and Vision CoT variants: VCoT$_{G_T}$(Tiny) and VCoT$_{G_L}$ (Large). Using mean Average Precision (mAP) as the evaluation metric for object detection, Table 3 demonstrates a positive correlation between VQA accuracy



(a) Performance Comparison in Open-end Question (b) Quantitative comparison analysis of Vision CoT. The darker parts, the more contributions.

Fig. 3: (a) A radar plot compares the performance of V2T-CoT and MedThink across three datasets: VQA-RAD(VR), SLAKE(SK), PathVQA(PV) using open-ended evaluation metrics. (b) A quantitative heatmap visualization showing that incorporating Vision CoT improves attention to the organ (liver).

and mAP. Results indicate that Vision CoT variants consistently enhance diagnostic performance, with higher mAP directly contributing to improved medical diagnosis accuracy. In Table 4, we conducted cross-region validation on five anatomical regions from SLAKE to evaluate the impact of organ-specific annotations in VCoT_{G_L} on VQA performance. Results show that mAP strongly correlates with VQA accuracy (Pearson's r=0.92), highlighting Vision CoT's ability to integrate anatomical-aware attention for improved diagnostic reasoning.

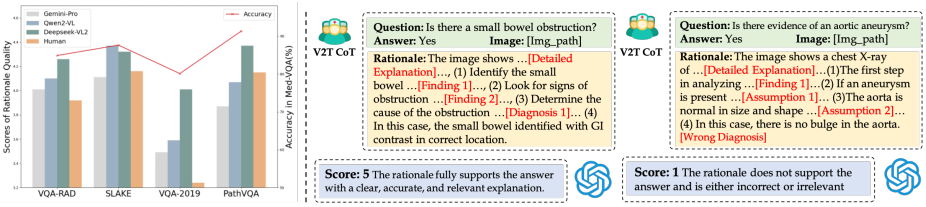
mAP(%)	RTM-Det 42.20	VCoT_{G_T} 51.70	VCoT_{G_L} 61.00
VQA-RAD	77.57	$80.88_{+3.31}$	$84.86_{+7.29}$
SLAKE	82.83	$83.41_{+0.58}$	$87.61_{+4.78}$
PathVQA	90.50	$90.83_{+0.33}$	$91.30_{+0.80}$
Med-2019	78.10	$78.10_{+0.00}$	$80.10_{+2.00}$

Table 3: Comparison of different Vision Detection method performance and the effect on Related Med-VQA.

Regions	mAP(%)	Acc(%)
Abdomen	62.43	84.43
Brain	69.05	88.00
Lung	65.33	91.60
Neck	42.23	68.75
Pelvic Cavity	51.74	68.42

Table 4: mAP and accuracy for different body regions on SLAKE dataset.

Text CoT Rationale Analysis. In Fig. 4, we assess four Med-VQA datasets through LLM-based automated scoring (1–5 scale) and human manual validation, focusing on rationale clarity in diagnostic reasoning. Results show VQA-RAD and SLAKE achieve average scores exceed 4, with comparable performance in VQA-2019 and PathVQA (Fig. 4A). The evaluation trend positively correlates with medical diagnosis accuracy, validating our assessment metric. Case studies (Fig. 4B) contrast a Score 5 rationale with explicit reasoning against a Score 1 case showing hallucinated assumptions and diagnostic errors. This demonstrates Text CoT's ability to improve reliability via interpretable diagnostic reasoning.



(a) Rationale Quality Score (b) Samples of Assessment on Rationale

Fig. 4: LLM and Human evaluation for rationale from **R-Med 39K**. (a) A bar chart illustrating rationale quality scores using different LLMs and human on four Med-VQA datasets, highlighting that higher rationale quality correlates with better Med-VQA performance. (b) Sample rationale quality assessments demonstrating Text CoT's role in medical diagnostics.

4 Conclusion

In this paper, we propose V2T-CoT, a novel multimodal reasoning framework that significantly enhances the accuracy and interpretability of Med-VQA by Vision and Text CoT reasoning. Vision CoT locates disease-related visual cues, while Text CoT, trained on the constructed R-Med 39K instruction-tuning dataset, provides reasoning paths to enhance accuracy, interpretability, and transparency in the medical reasoning and diagnosis.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (Grant No. 62476241), the Natural Science Foundation of Zhejiang Province, China (Grant No. LZ23F020008), and the Zhejiang University-Angelalign Inc. R&D Center for Intelligent Healthcare.

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

References

1. Bazi, Y., Rahhal, M.M.A., Bashmal, L., Zuair, M.: Vision–language model for visual question answering in medical imagery. Bioengineering **10**(3), 380 (2023)
2. Bellagente, M., Tow, J., Mahan, D., Phung, D., Zhuravinskyi, M., Adithyan, R., Baicoianu, J., Brooks, B., Cooper, N., Datta, A., Lee, M., Mostaque, E., Pieler, M., Pinnaparju, N., Rocha, P., Saini, H., Teufel, H., Zanichelli, N., Riquelme, C.: Stable lm 2 1.6b technical report (2024)
3. Ben Abacha, A., Hasan, S.A., Datla, V.V., Demner-Fushman, D., Müller, H.: Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In: Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes. 9-12 September 2019 (2019)
4. Chen, J., Cai, Z., Ji, K., Wang, X., Liu, W., Wang, R., Hou, J., Wang, B.: Huatuogpt-o1, towards medical complex reasoning with llms (2024)

5. Chen, Z., Li, G., Wan, X.: Align, reason and learn: Enhancing medical vision-and-language pre-training with knowledge. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 5152–5161 (2022)
6. Eslami, S., Meinel, C., De Melo, G.: Pubmedclip: How much does clip benefit visual question answering in the medical domain? In: Findings of the Association for Computational Linguistics: EACL 2023. pp. 1181–1193 (2023)
7. Gai, X., Zhou, C., Liu, J., Feng, Y., Wu, J., Liu, Z.: Medthink: Explaining medical visual question answering via multimodal decision-making rationale. arXiv preprint arXiv:2404.12372 (2024)
8. Gong, H., Chen, G., Liu, S., Yu, Y., Li, G.: Cross-modal self-attention with multi-task pre-training for medical visual question answering. In: Proceedings of the 2021 International Conference on Multimedia Retrieval. pp. 456–460 (2021)
9. Gunasekar, S., Zhang, Y., Aneja, J., Mendes, C.C.T., Giorno, A.D., Gopi, S., Javaheripi, M., Kauffmann, P., de Rosa, G., Saarikivi, O., Salim, A., Shah, S., Behl, H.S., Wang, X., Bubeck, S., Eldan, R., Kalai, A.T., Lee, Y.T., Li, Y.: Textbooks are all you need (2023)
10. He, J., Li, P., Liu, G., Zhao, Z., Zhong, S.: Pefomed: Parameter efficient fine-tuning on multimodal large language models for medical visual question answering. arXiv preprint arXiv:2401.02797 (2024)
11. He, X., Zhang, Y., Mou, L., Xing, E., Xie, P.: Pathvqa: 30000+ questions for medical visual question answering. arXiv preprint arXiv:2003.10286 (2020)
12. Jiang, S., Wang, Y., Song, S., Zhang, Y., Meng, Z., Lei, B., Wu, J., Sun, J., Liu, Z.: Omniv-med: Scaling medical vision-language model for universal visual understanding (2025), https://arxiv.org/abs/2504.14692
13. Jiang, S., Zheng, T., Zhang, Y., Jin, Y., Yuan, L., Liu, Z.: Med-moe: Mixture of domain-specific experts for lightweight medical vision-language models. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 3843–3860 (2024)
14. Khare, Y., Bagal, V., Mathew, M., Devi, A., Priyakumar, U.D., Jawahar, C.: Mmbert: multimodal bert pretraining for improved medical vqa. In: 2021 IEEE 18th International Symposium on Biomedical Imaging. pp. 1033–1036 (2021)
15. 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**(1), 1–10 (2018)
16. 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. Advances in Neural Information Processing Systems **36** (2024)
17. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., Chang, K.W., Gao, J.: Grounded language-image pre-training. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10955–10965 (2022)
18. 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. pp. 1650–1654 (2021)
19. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36** (2024)
20. Liu, J., Hu, T., Zhang, Y., Feng, Y., Hao, J., Lv, J., Liu, Z.: Parameter-efficient transfer learning for medical visual question answering. IEEE Transactions on Emerging Topics in Computational Intelligence (2023)

21. Liu, J., Wang, Y., Du, J., Zhou, J.T., Liu, Z.: MedCoT: Medical chain of thought via hierarchical expert. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 17371–17389 (2024)
22. Liu, X., Liu, H., Yang, G., Jiang, Z., Cui, S., Zhang, Z., Wang, H., Tao, L., Sun, Y., Song, Z., et al.: A generalist medical language model for disease diagnosis assistance. Nature Medicine pp. 1–11 (2025)
23. Liu, Y., Wang, Z., Xu, D., Zhou, L.: Q2atransformer: Improving medical vqa via an answer querying decoder. In: International Conference on Information Processing in Medical Imaging. pp. 445–456. Springer (2023)
24. Qi, Z., Fang, Y., Zhang, M., Sun, Z., Wu, T., Liu, Z., Lin, D., Wang, J., Zhao, H.: Gemini vs gpt-4v: A preliminary comparison and combination of vision-language models through qualitative cases. arXiv preprint arXiv:2312.15011 (2023)
25. Savage, T., Nayak, A., Gallo, R., Rangan, E., Chen, J.H.: Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. NPJ Digital Medicine **7**(1), 20 (2024)
26. Van Sonsbeek, T., Derakhshani, M.M., Najdenkoska, I., Snoek, C.G., Worring, M.: Open-ended medical visual question answering through prefix tuning of language models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 726–736. Springer (2023)
27. Xiong, H., Wang, S., Zhu, Y., Zhao, Z., Liu, Y., Huang, L., Wang, Q., Shen, D.: Doctorglm: Fine-tuning your chinese doctor is not a herculean task. arXiv preprint arXiv:2304.01097 (2023)
28. Zhang, K., Yu, J., Adhikarla, E., Zhou, R., Yan, Z., Liu, Y., Liu, Z., He, L., Davison, B., Li, X., et al.: Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. arXiv e-prints pp. arXiv–2305 (2023)
29. Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., Smola, A.: Multimodal chain-of-thought reasoning in language models. arXiv preprint arXiv:2302.00923 (2023)