MedVCTP: Improving Accuracy and Explainability in Medical Visual Reasoning

Aman Syed[2*+] **Siwon Ryu**[1] **Nayan Saxena**[2+] **Kevin Zhu**[1]

Abstract

Reasoning transparency and accuracy are critical to the implementation of AI algorithms in medical applications. However, modern medical visual-language models (VLMs) often generate conclusions without explicit reasoning, limiting clinician trust and potentially compromising the quality of diagnosis. Reasoning-focused VLMs remain confined to basic VQA datasets (e.g., A-OKVQA), while medical VLMs lack reasoning transparency, modularity, and output refinement capabilities. We introduce Medical Visual Chain-of-Thought Processing (MedVCTP), a training-free framework implementing a structured See–Think–Confirm pipeline. The See stage extracts global and regional visual concepts via advanced visual encoders. The Think stage generates reasoning-grounded answers through LLM-based chain-of-thought processing. The Confirm stage iteratively refines rationales via multi-shot prompting and cross-modal CLIP-based consistency checks, aligning reasoning with visual context to mitigate hallucination and enable visual grounding. Our modular design supports rapid deployment with interchangeable components for scalable performance. On SLAKE, MedVCTP achieves 85.8% accuracy-a 19.4% improvement over ablations without CLIP refinement—demonstrating that iterative cross-modal validation directly enhances both accuracy and reasoning coherence. These results establish MedVCTP as a step toward reliable, explainable medical visual reasoning systems deployable without task-specific training. Code and artifacts are available at https://github.com/Carrote-s/MedVCTP.

1 Introduction

Reasoning transparency and accuracy are critical to the implementation of AI algorithms in medical applications [26]. However, modern medical visual-language models (VLMs) often generate conclusions without explicit reasoning [23], limiting clinician trust and potentially compromising the quality of diagnosis. To overcome these limitations, transparent reasoning is crucial in AI, as it explains the logic behind a decision and the boundaries that govern its conclusions, allowing healthcare professionals to examine predictions with greater care [19]. Moreover, powerful reasoning-focused VLMs in broader domains tend to focus on basic VQA datasets, making them difficult to adapt to specialized tasks such as medical visual question answering (Med-VQA) [31]. Additionally, high stake fields such as medicine require medical conclusion refinement to minimize visual hallucination, which is also a gap in many medical VLMs.

Recent work has attempted to address these gaps. Since reasoning and chain-of-thought prompting are known to improve complex problem solving [35], Visual Chain-of-Thought Prompting (VCTP) was introduced to improve rationale transparency and accuracy in VQA [5]. Although VCTP demonstrates the potential of CoT reasoning, we argue that its focus on standard VQA datasets [5] limits its transferability to various complex domains, such as medicine. Alternatively, reinforcement learning has been applied to improve reasoning quality in visual language medical models [23].

[1]Algoverse AI Research, [2]Independent, [*]Lead Author, [+]Project Lead ✉ : amansyedcs@gmail.com

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

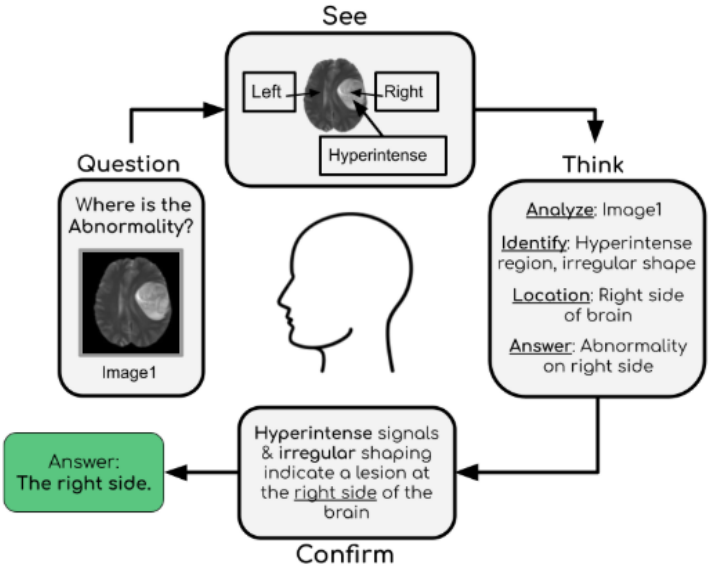

Figure 1: Overview of the MedVCTP (Medical Visual Chain-of-Thought Processing) framework. The diagram illustrates the three-stage See–Think–Confirm reasoning pipeline, where visual features are extracted, interpreted via chain-of-thought prompting, and refined through cross-modal verification.

However, we note that these methods have several limitations: (1) They rarely incorporate chain-of-thought reasoning or iterative refinement loops; (2) Their reliance on narrow visual encoding limits their ability to generalize across modalities; and (3) They are difficult to refine due to a lack of modularity. Approaches such as MC-CoT [36] have shown the benefits of structured reasoning in Med-VQA, significantly improving zero-shot performance by integrating multimodal medical knowledge. A large gap in such papers is their lack of addressing hallucination, where a lack of diagnosis refinement or grounding in visual images brings space in for inaccurate answers and hallucination.

To bridge these gaps, we propose MedVCTP, a modular, visually grounded reasoning pipeline adapted from the VCTP framework [5]. MedVCTP splits the Med-VQA process into three modules, See, Think, and Confirm, enabling interpretable reasoning that leverages both visual and textual cues. This design improves reasoning quality and transparency, supports iterative refinement, and boasts a modular design for medical applications. MedVCTP integrates visual and textual cues into a modular reasoning framework, allowing iterative refinement of intermediate conclusions and direct grounding of rationale in observed anatomical features. MedVCTP is one of the first pipelines to directly utilize medical CoT prompting not only to generate rationales, but to attend to important visual concepts for reasoning. Additionally, a central objective of MedVCTP is to provide not only transparent, but also accurate rationales. By utilizing similarity scores to compare generated rationales with respect to images, MedVCTP iteratively refines such rationales to mitigate hallucination and improve visual grounding of rationales in such images. Moreover, MedVCTP boasts high generalizability by using state-of-the-art visual encoders [20] and prompting methods [3] that allow for dataset specific tuning.

As illustrated in Figure 1, MedVCTP mirrors the reasoning process clinicians follow [6]. First, the model identifies and characterizes anatomical structures and their relationships, including structures such as organs and tumors, and various characteristics such as shape, size, image quality, plane, and other features. After analyzing the question type to determine the required information, the model selects the most relevant concepts and relationships to generate a final rationale that integrates these observations into a coherent, interpretable answer. In our pipeline, we add a loop at the end that iterates through the Think and Confirm stages, based on a score the model assigns to a rationale to determine how visually grounded it is. This score is a CLIP similarity score, measuring the similarity between the image and the rationale to prevent hallucination [27]. By doing so, MedVCTP ultimately

generates rationales that are visually grounded in medical images, that stay on topic with respect to both the question and the image.

For example, when answering the question, "Where is the abnormality?" (Figure 1), the pipeline first identifies the brain as the primary organ and detects a hyperintense region on the right side, characterized by irregular shape and bright signal intensity. It then determines that the question focuses on detecting an abnormality and its location, causing the model to prioritize relevant positional and pathological information. The model ultimately concludes that the abnormality is on the right side of the brain, based on the hyperintense signals and irregular shape observed in that region. This step-by-step reasoning mirrors how clinicians approach similar tasks [6] and provides transparent evidence supporting the conclusion.

Contributions.

- **Modular, training-free Med-VQA with visual grounding.** We adapt and extend a See–Think–Confirm pipeline to clinical imaging, coupling global *and* region-level captions with chain-of-thought prompting. The design is inference-only and component-swappable (encoders / LLMs / verifiers), and works with or without bounding boxes.
- **Iterative cross-modal verification to curb hallucination.** A BiomedCLIP-based cosine-similarity check scores image–rationale alignment and triggers revision until a grounding threshold or iteration cap is reached, yielding interpretable, clinically relevant rationales.
- **Competitive accuracy with lightweight models + reproducibility.** On SLAKE close-ended questions, MedVCTP attains **85.8%** accuracy using Llama 3.1 8B and MedGemma, **+19.4** points over an ablation without verification. We include anonymized prompts, few-shot exemplars, and implementation details for replication.

MedVCTP does these while maintaining modularity for domain adaptation.

2 Related Works

2.1 Medical VQA

PMC-VQA and Generalist Model Baselines PMC-VQA [43] reframes medical VQA as a generative task, aligning a vision encoder with a language model and introducing a 227k-pair dataset. It outperforms generalist models like Open-Flamingo [2] and BLIP-2 [13], reaching 78% on VQA-RAD [12]. However, it lacks explicit reasoning (e.g., chain-of-thought), limiting interpretability.

Reinforcement Learning–Driven Reasoning MedVLM-R1 [23] enhances reasoning in medical VLMs using the GRPO reinforcement learning method [30] and supervised fine-tuning. Despite its compact 2B-parameter size and training on only 600 MRI samples, it achieves 95.3% in-domain accuracy and 70% on CT and X-ray, outperforming much larger models. RL encourages more structured, clinically meaningful outputs, but limited modality exposure and lack of chain-of-thought or refinement loops restrict generalization. Similar approaches, such as Med-R1 [11] and RARL [25], improve reasoning via RL but treat it as a single, non-modular process.

Models Evaluated on the SLAKE Dataset Recent work in medical VQA has explored a variety of architectures. LLaDA-MedV [7] is a fine-tuned diffusion model that uses an iterative de-masking process to generate responses. M212 [15] is another fine-tuned model that utilizes self-supervised learning for VQA. Other traditional approaches, such as VGGseg+SAN and VGG+SAN [18], utilize a VGG-19 CNN [18] for image feature extraction and a Stacked Attention Network (SAN [38]) to answer questions, with VGGseg+SAN incorporating image segmentation. Unlike these models, MedVCTP focuses on a modular reasoning framework.

2.2 Improving Chain-of-Thought in VQA

Chain-of-thought (CoT) prompting improves reasoning in LLMs by decomposing complex problems into intermediate steps, as shown by [35] with PaLM 540B. Extending this to VQA, recent work integrates CoT with visual grounding to enhance interpretability and accuracy. Visual Chain-of-Thought Prompting (VCTP) [5] introduces a modular pipeline—see, think, confirm—that combines

visual grounding (VinVL [40], BLIP [14]), reasoning with LLMs (OPT [42], Llama [34]), and verification via CLIP [27]. Iterating this process yields higher accuracy and more coherent rationales, outperforming baselines like BLIP-2 [13] and PICa [37] on A-OKVQA [28] and OK-VQA [21]. While VCTP validates the value of modular CoT in VQA, it has been tested mainly on general datasets, not medical domains.

2.3 Image captioning

BLIP-2 Image captioning, born from linking computer vision and NLP, has driven advances in vision–language models. BLIP-2 bridges the modality gap by combining frozen image encoders and large language models with a lightweight Querying Transformer [13]. Using a two-stage pretraining strategy, it outperforms larger models like Flamingo-80B [1] on VQA and captioning tasks with far fewer trainable parameters. While effective for general applications, BLIP-2 was not designed for medical use, a gap later addressed by MedBLIP [4].

Medical Image Captioners Medical image captioning (MIC) requires clinically precise and semantically rich descriptions, making general-purpose VLMs like BLIP [14], BLIP-2 [13], Gemini [32], and ViT-GPT2 [10] prone to generic or inaccurate outputs. Fine-tuning on domain-specific datasets such as ROCO [24] improves radiology captioning, with BLIP-2 variants and ViT-GPT2 showing better accuracy and interpretability after adaptation [16]. However, such fine-tuning often limits generalizability. MedGemma (4B, 27B) [29], built on the Gemma 3 architecture [33] with a medically tuned SigLIP encoder [39] (MedSigLIP), addresses this by combining general multimodal strength with specialized medical understanding, advancing performance in both VQA and image captioning.

2.4 BiomedCLIP

Visual grounding techniques originated from CLIP models, which compute similarity scores between text and images. BiomedCLIP [41] adapts this approach for medical applications, introducing PMC-15M [43], a diverse dataset of 15 million image-text pairs. Using a pre-trained BiomedCLIP model, they achieve state-of-the-art results for biomedical imaging tasks including VQA, concluding that large-scale pretraining on diverse data is a highly effective method for creating a generalist biomedical model. This serves as the main cross modal checker for rationale refinement in MedVCTP.

3 Methodology

In this section, we present our novel Med-Visual Chain-of-Thought Prompting (MedVCTP) model, a comprehensive and robust framework for medical VQA. MedVCTP utilizes a powerful visual encoder [29] for visual understanding of medical images, a sophisticated reasoning LLM [34] with specialized prompting functions depicted in Figure 2, guided by a chain-of-thought process, and a BiomedCLIP model [41] to do a cross verification to prevent hallucination and refine reasoning.

The model is broken up into three distinct and sequential stages: the See, Think, and Confirm modules. In the See module, the image and its bounding boxes go through a medical VLM, such as MedGemma [29], to create regional captions for each bounding box. The VLM then generates a global caption based off the entire raw image, providing a high-level summary of its content. These captions are generated with specific prompting instructions to guide coherency and accuracy.

See Module The See module converts raw medical images into structured textual representations for downstream reasoning. Using MedGemma [29], each image generates a global caption describing the full image and regional captions for every bounding box. For regional captions, MedGemma receives both the cropped region and the full image as context, along with the region label, ensuring the LLM [9] can capture key features while retaining broader clinical context. This dual-input approach prevents MedGemma from misinterpreting isolated crops. Utilizing regional captioning is optional, as the framework's modular design allows for quick switches on captioning type. All generated features are indexed by image ID to enable efficient retrieval in the Think module.

Think Module The Think module serves as MedVCTP's core reasoning engine. Llama 3.1 Instruct 8B [9], chosen for efficient instruction-following and reasoning, receives the question along with

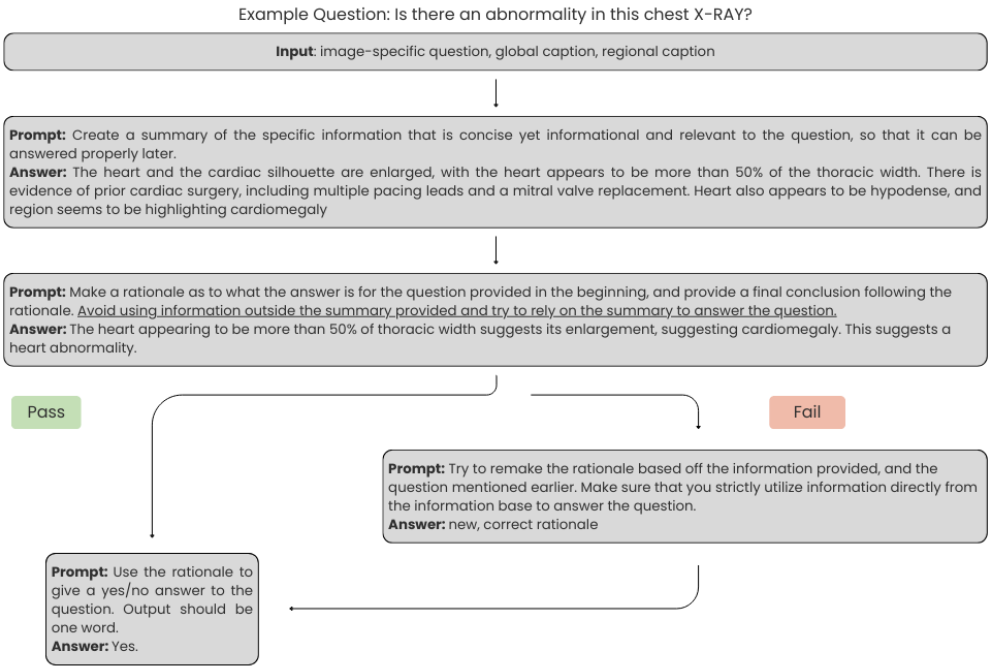

Figure 2: Prompt flowchart used in the MedVCTP Think and Confirm stages. It shows how question–image pairs are transformed into structured prompts for the large-language-model reasoning process. Example prompts are abbreviated versions of those used in implementation.

the image ID to retrieve relevant global and regional captions from the See module. A specialized prompting function guides the LLM to focus on pertinent visual information, filtering out irrelevant context and reducing hallucination. This relevance-focused reasoning ensures that downstream outputs remain visually grounded and accurate.

Confirm Module The Confirm module performs final reasoning and cross-modal verification. Precomputed features from the Think stage are fed to a prompting function with 5–7 exemplars; we decode with temperature 0.3. We then encode the image and the generated rationale with BiomedCLIP, L2-normalize both embeddings, and compute their cosine similarity $s = \hat{\mathbf{v}}_{\text{img}}^{\top} \hat{\mathbf{v}}_{\text{txt}}$. A rationale is accepted if $s \geq \tau$ (empirically chosen); otherwise the LLM is prompted to revise, up to $K = 3$ iterations with early stopping on acceptance. This gating reduces hallucination and yields more visually grounded, clinically relevant rationales.

Looping Mechanism Using BiomedCLIP [41], a similarity score is computed between the rationale and its associated image and compared to a threshold optimized for medical VQA through the process. If the score meets the threshold, the LLM generates a conclusion from the rationale and question. If not, the LLM receives feedback about insufficient visual grounding, the previous rationale, the similarity score, and context, and is instructed to regenerate the rationale and conclusion. This process repeats until the threshold is met or maximum iterations are reached, safeguarding against hallucination and ensuring outputs remain visually grounded and medically accurate.

4 Experiment

In this section we demonstrate the modularity and reasoning of our approach through accuracy and an ablation study. We perform the main MedVCTP experiment along with the ablation skipping cross modal checks. All experiments were performed on a single NVIDIA B200 GPU with 180 GB of VRAM, provided through the RunPod platform. Our implementation was developed using the PyTorch framework, leveraging its capabilities for efficient model management and inference. All

Table 1: Qualitative examples showing MedVCTP's step-by-step reasoning on real clinical VQA cases. Each row lists the question, the model's generated rationale, and the predicted answer, giving users insights into visually grounded reasoning.

Question	Model Rationale (MedVCTP's Output)	Predicted Answer
Is the lung abnormal?	A rounded opacity in the left mid-lung field, measuring approximately 4 cm in diameter, suggests a potential lung abnormality, such as a mass or tumor. The opacity's appearance and location in the periphery of the left lung support this conclusion.	**Yes**
Are there abnormalities in this image?	The chest X-ray image shows an enlarged heart, which is predominantly hypodense, suggesting a possible abnormality. The heart's enlarged size is likely related to prior cardiac surgery, including multiple pacing leads and a mitral valve replacement.	**Yes**
Is this image taken via MRI?	The image is described as a T1-weighted axial MRI image of the brain, which directly indicates that it was taken via MRI. This is further supported by the imaging characteristics and features mentioned, such as hypointense areas and increased signal intensity.	**Yes**

prompt templates, prompt engineering details, and configuration specifics for MedVCTP are available in our GitHub repository.

4.1 Model Details

Open source models are strictly used. We utilize MedGemma 4B instruction-tuned [29] for global and regional caption generation. Llama 3.1 8B Instruct [9] is selected for concept attendance and conclusion generation, given its faster inference while maintaining reasoning and accuracy. We use a max iteration of 3 for our main pipeline and 5 prompts with rationales and conclusions to guide LLM response and accuracy. Information about caption generation for MedGemma and LLama 3.1 8B is provided in the repository.

4.2 Dataset and Evaluation Metrics

We test our model on the SLAKE benchmark [17], which provides VQA questions and bounding box annotations. The total dataset contains 7033 questions excluding Chinese questions, 2167 of which are close-ended questions with bounding boxes (Yes/No answers). SLAKE provides bounding box annotations along with raw images and questions for assessing VQA model accuracy. The main preprocessing steps utilized are isolating the close-ended questions with bounding boxes to create our main experimental dataset to assess performance. We solely test our model on accuracy for the close-ended bounding box questions, which are such 2167 questions. We choose SLAKE because of its high generalization to various datasets due to its coverage of 10 different question types, multiple modalities, and contains images of many parts of the body. Accuracy is calculated by simply dividing number of correct answers by total questions answered.

4.3 Baselines

We compare our enhanced reasoning VLM to other models evaluated on SLAKE performance, including LLaDA-MedV [7], VGGseg + SAN [18], VGG + SAN [18], M212 [15] and PubMedCLIP [8]. The implementation of an 8B parameter LLM as the core for MedVCTP demonstrates the

Table 2: SLAKE close-ended (with boxes). Accuracy is %. Baselines are reported from their papers; MedVCTP is inference-only (no fine-tuning).

Model	Accuracy	Notes
MedVCTP (ours)	85.8	With CLIP grounding; no fine-tuning
Ablation	66.4	Without CLIP grounding; no fine-tuning
LLaDA-MedV [7]	93.21	Diffusion-based VLM; fine-tuned
VGGseg+SAN [18]	79.84	VGG-19 + segmentation; fine-tuned
VGG+SAN [18]	76.13	VGG-19; fine-tuned
M212 [15]	91.10	Self-supervised; fine-tuned
PubMedCLIP [8]	82.50	CLIP; fine-tuned

accessibility and scalability of the pipeline for future purposes. We create an ablation study by removing the BiomedCLIP [41] iterative cross-modal check and setting this variant as our baseline for comparison to our main experiment implementing BiomedCLIP, a relatively quick process given MedVCTP's modularity. This is for the purpose of demonstrating the potential performance increase on medical VQA when utilizing CLIP models for visual grounding. These baseline models aren't run directly by us, as results are extracted from respective research papers. We only run the MedVCTP experiment along with the ablation baseline.

5 Results

Our proposed MedVCTP model achieves 85.8% accuracy on the SLAKE dataset [17], representing a substantial 19.4% improvement over a baseline ablation study that excludes the cross-modal verification loop. Notably, our approach employs Llama 3.1 8B Instruct model [9] without specific fine-tuning, leveraging the MedVCTP pipeline's reasoning capabilities to achieve competitive performance. Our model outperforms several established approaches on SLAKE, including PubMedCLIP [8], VGGseg + SAN [18], and VGG + SAN [18], demonstrating the effectiveness of reasoning-based medical VLMs over traditional non-reasoning architectures. However, specialized medical VLMs including LLaDA-MedV [7] and M212 [15] achieve superior performance, exceeding our model by 6.51% and 5.3% respectively. This suggests that more fine-tuned and powerful state-of-the-art models outperform our approach that uses a general LLM. LLaDA-MedV is a fine-tuned diffusion model using an iterative de-masking process to generate more coherent and contextually rich responses by filling in masked parts of text based on both the user's question and the visual features from the image. The remaining performance gap to perfect accuracy underscores the inherent limitations of employing a general-purpose 8B parameter model, despite instruction tuning, in complex medical reasoning tasks requiring both visual understanding and domain expertise.

Table 1 shows three examples with the question, generated rationale, and predicted answer. This table demonstrates the exhaustive reasoning process of the model, improving accuracy and interpretability.

5.1 Ablation

For our ablation study, we tested a baseline version of MedVCTP that excludes the BiomedCLIP [41] refinement loop. This variant uses the first-pass rationale as its final answer, omitting the iterative cross-modal check. We set the example prompt for few-shot prompting to 5 examples similar to the original experiment, and used MedGemma [29] for caption extraction, Llama 3.1 8B Instruct [9] for concept attendance and conclusion generation, identical to the main experiment. This ablation resulted in a performance drop of 19.4%, indicating the importance of the cross-modality check.

5.2 Future Works and Limitations

MedVCTP is currently evaluated using lower capacity LLMs and captioners, which doesn't fully capture its potential in medical VQA. Additionally, MedVCTP was evaluated on a single dataset SLAKE, as SLAKE [17] is a large yet exhaustive VQA dataset, providing large amounts of questions for analysis. In the future, MedVCTP can be evaluated on multiple datasets. Although more powerful reasoning models such as GPT-4o [22] exist, we did not utilize them due to prohibitive API costs.

Furthermore, open-ended accuracy was not assessed due to time and resource constraints, which limits the scope of our findings. Similarly, due to computational costs and the large expanse of SLAKE VQA we were unable to run multiple independent evaluations.

Another limitation is that the subset of closed-ended questions extracted from SLAKE may have included knowledge-graph-based questions. Our pipeline did not use the provided knowledge graph triplets, which may have hindered performance on these specific questions. Lastly, our evaluation excluded 227 of the 2394 closed-ended questions, as they referred to images where no bounding box annotations were provided. Although this is the case, our modular design allows for quick modification to remove the bounding box necessity, to adapt to more datasets. In the future, this limitation could also be addressed by developing methods to automatically generate bounding boxes and labels, thereby proposing another solution to improving the model's generalization. However, this direction requires substantial further research.

6 Conclusion

In this paper, we introduced a novel approach to medical VQA, MedVCTP, which decomposes the VQA process into a modular See, Think, and Confirm framework. Through an ablation study, we demonstrated the importance of our cross-modal refinement module and showcased the approach's ability to facilitate complex chain-of-thought reasoning for medical scenarios. While our model achieves accuracy superior to that of general-purpose methods, it is outperformed by larger, state-of-the-art, specialized medical VLMs. However, MedVCTP's modular design allows it to be easily modified and improved with more powerful LLMs and captioners. MedVCTP's modular design facilitates future integration with advanced models and datasets, paving the way for improved, interpretable medical visual reasoning with reduced hallucination and transparent reasoning.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[2] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[4] Qiuhui Chen and Yi Hong. Medblip: Bootstrapping language-image pre-training from 3d medical images and texts. In *Proceedings of the Asian conference on computer vision*, pages 2404–2420, 2024.

[5] Zhenfang Chen, Qinhong Zhou, Yikang Shen, Yining Hong, Zhiqing Sun, Dan Gutfreund, and Chuang Gan. Visual chain-of-thought prompting for knowledge-based visual reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1254–1262, 2024.

[6] Pat Croskerry. A universal model of diagnostic reasoning. *Academic medicine*, 84(8):1022–1028, 2009.

[7] Xuanzhao Dong, Wenhui Zhu, Xiwen Chen, Zhipeng Wang, Peijie Qiu, Shao Tang, Xin Li, and Yalin Wang. Llada-medv: Exploring large language diffusion models for biomedical image understanding. 2025.

[8] Sedigheh Eslami, Christoph Meinel, and Gerard de Melo. PubMedCLIP: How much does CLIP benefit visual question answering in the medical domain? In Andreas Vlachos and Isabelle Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1181–1193, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[9] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[10] Md Rakibul Islam, Md Zahid Hossain, Mustofa Ahmed, Most Samu, et al. Vision-language models for automated chest x-ray interpretation: Leveraging vit and gpt-2. *arXiv preprint arXiv:2501.12356*, 2025.

[11] Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, and Xiaofeng Yang. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *arXiv preprint arXiv:2503.13939*, 2025.

[12] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.

[13] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[14] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[15] Pengfei Li, Gang Liu, Lin Tan, Jinying Liao, and Shenjun Zhong. Self-supervised vision-language pretraining for medial visual question answering. *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2022.

[16] Manshi Limbu and Diwita Banerjee. Medblip: Fine-tuning blip for medical image captioning, 2025.

[17] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1650–1654. IEEE, 2021.

[18] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Fang Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654, 2021.

[19] Yushu Liu, Chenxi Liu, Jianing Zheng, Chang Xu, and Dan Wang. Improving explainability and integrability of medical ai to promote health care professional acceptance and use: Mixed systematic review. *Journal of Medical Internet Research*, 27:e73374, 2025.

[20] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[21] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019.

[22] OpenAI. Gpt-4v(ision) system card. `https://openai.com/research/gpt-4v-system-card`, 2023.

[23] Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. *arXiv preprint arXiv:2502.19634*, 2025.

[24] Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and C. Friedrich. Radiology objects in context (roco): A multimodal image dataset. In *CVII-STENT/LABELS@MICCAI*, 2018.

[25] Tan-Hanh Pham and Chris Ngo. Rarl: Improving medical vlm reasoning and generalization with reinforcement learning and lora under data and hardware constraints. *arXiv preprint arXiv:2506.06600*, 2025.

[26] Thomas P Quinn, Stephan Jacobs, Manisha Senadeera, Vuong Le, and Simon Coghlan. The three ghosts of medical ai: Can the black-box present deliver? *Artificial intelligence in medicine*, 124:102158, 2022.

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[28] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer, 2022.

[29] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

[30] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[31] Violetta Shevchenko, Damien Teney, Anthony Dick, and Anton van den Hengel. Reasoning over vision and language: Exploring the benefits of supplemental knowledge. *arXiv preprint arXiv:2101.06013*, 2021.

[32] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[33] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[34] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[35] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[36] Lai Wei, Wenkai Wang, Xiaoyu Shen, Yu Xie, Zhihao Fan, Xiaojin Zhang, Zhongyu Wei, and Wei Chen. Mc-cot: A modular collaborative cot framework for zero-shot medical-vqa with llm and mllm integration. *arXiv preprint arXiv:2410.04521*, 2024.

[37] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 3081–3089, 2022.

[38] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21–29, 2015.

[39] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.

[40] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588, 2021.

[41] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

[42] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

[43] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.