MedVCoT: Bridging the Modality Gap in Medical VQA Through Latent Visual Reasoning

Bo Xu[1] , **Quanhao Zhu**[1] , **Boling Zhu**[1] , **Chenyuan Wang**[1] , **Liang Zhao**[1*] , **Hongfei Lin**[1] and **Feng Xia**[2]
[1]Dalian University of Technology, China
[2]RMIT University, Australia
{boxu, liangzhao, hflin}@dlut.edu.cn, zhuqh19@gmail.com, SinsapZ@outlook.com, chenyuanwang@mail.dlut.edu.cn, f.xia@ieee.org

Abstract

With the rising demand for trustworthy AI in clinical practice, strong interpretability is now a critical requirement as well as accuracy. However, the modality gap for medical visual question answering is quite severe when continuous visual signals are forcibly projected into discrete text space for reasoning, and the loss of necessary diagnostic information leads to low precision and black-box opacity. To address this problem, we propose MedVCoT, which incorporates latent visual reasoning into the medical visual question answering(VQA) domain. Rather than merely integrating modules, MedVCoT utilizes the specialized expertise of MedSAM to train a large vision-language model so that it can autonomously generate consistent and continuous latent visual tokens within Visual Chain-of-Thought. This mechanism forces the model to explicitly "see" the lesion in the latent space before formulating a textual diagnosis, ensuring answers are causally rooted in verifiable visual evidence rather than statistical hallucination. We achieve this through a progressive 3-stage training procedure: medical feature alignment, visual reasoning learning by utilizing latent tokens generated, and instruction tuning for complex clinical scenarios. Extensive experiments show that MedVCoT can achieve state-of-the-art performance on multiple benchmarks, outperforming other methods by large margins. Meanwhile, it provides pixel-level segmentation masks to validate its diagnostic reasoning. Our demo is available at https://zhuqh19.github.io/MedVCoT.

1 Introduction

Medical Visual Question Answering (Medical VQA) stands as a cornerstone of AI-assisted diagnosis, which needs both high accuracy and rigorous interpretability. The "black-box" diagnosis made to a patient, although it may be correct, will be unconvincing without verifiable visual evidence. However, current state-of-the-art Medical Vision-Language Models (VLMs) come to a fundamental bottleneck: the modality gap. When continuous and fine-grained visual signals (such as the boundaries of a lesion or texture gradients) are forcibly projected into a discrete textual space for reasoning, there is a lot of visual evidence ignored inevitably. This information loss is particularly detrimental in clinical practice, where the subtlest cues in visual content can matter to diagnosis. In that connection, the lack of direct visualization of reasoning has been a barrier to the application of medical AI systems applied to clinical scenarios.

Prior attempts to address this challenge fall short in large part because they fail to explicitly ground reasoning on visual information. Early medical VLMs, such as LLaVA-Med [Liu *et al.*, 2023; Li *et al.*, 2023], mainly focused on aligning the global image features with text through large-scale instruction tuning. However, although such models can generate fluent general descriptions, they are fundamentally black box in nature, revealing no information about which particular regions of an image or features led to a particular diagnosis [Mullappilly *et al.*, 2024; Chen *et al.*, 2024; Ning *et al.*, 2025; He *et al.*, 2024; Xu *et al.*, 2025c; Xu *et al.*, 2025a]. This uninterpretability is extremely critical in a clinical scenario, where objective evidence is everything. Other works then attempted to introduce Chain-of-Thought (CoT) reasoning [Wei *et al.*, 2022] to improve interpretability [Hong *et al.*, 2025; Liu *et al.*, 2024; Servantez *et al.*, 2024; Li *et al.*, 2025]. However, these text-based CoTs are limited by the modality gap: since complex, continuous visual information, such as the exact contours of a tumor or subtle texture variations, is forcibly projected into discrete textual descriptions (e.g., merely stating irregular shape), they fundamentally lead to lossy representations of data and therefore also lose precision. Thus, neither paradigm succeeds in providing the pixel-level grounding necessary for high-stakes medical decision-making and so leaves the interpretability gap largely unbridged and clinicians without a reliable way to verify the model's focus.

To address these challenges, we propose MedVCoT, the first approach to incorporate a Visual Chain-of-Thought (Visual CoT) into medical VQA. Different from traditional methods, which only perform reasoning in the text space, our model generates a sequence of continuous latent visual tokens autonomously within a separate thought horizon (denoted as `<think>` tags) before answering. These tokens

*Corresponding author.

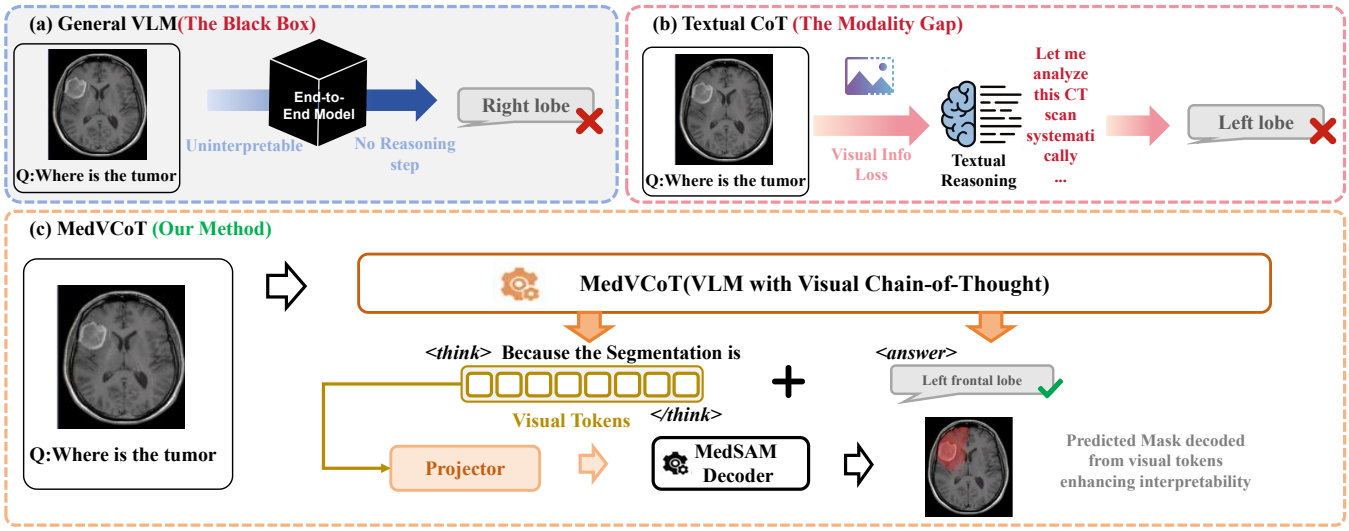

Figure 1: Comparison of reasoning paradigms in Medical VQA. (a) General VLMs operate as uninterpretable black boxes. (b) Text-only CoT suffers from visual information loss due to the modality gap. (c) Our MedVCoT incorporates a Visual Chain-of-Thought , ensuring strong interpretability and high accuracy.

act as a bridge between linguistic reasoning and visual perception. In particular, under the guidance of the specialized medical visual expert MedSAM, the VLM learns to generate these tokens as an intermediate reasoning output. These generated tokens can be further mapped with a projector to trigger MedSAM to decode explicit segmentation masks. This mechanism forces the model to "see" and localize the lesion in latent space before "speaking" the diagnosis so as to ensure that answers are causally predicated on verifiable visual evidence.

Our contributions are threefold:

- **Pioneering Framework for Latent Visual Reasoning in Medical VQA**: We propose MedVCoT, the first medical-VQA-specific framework that performs reasoning through continuous latent visual embeddings instead of discrete text tokens. Through incorporating a Visual Chain-of-Thought mechanism, MedVCoT enables the model to carry out structured visual introspection within a latent thought horizon, ultimately bridging the modality gap that constrains conventional vision-language models.
- **Verifiable Interpretability with Latent-to-Mask Projection Projection**: Contrary to black-box systems, MedVCoT generates verifiable visual evidence for interpretability in addition to textual diagnoses. The generated latent visual tokens of VLM during reasoning can be projected to trigger MedSAM for decoding pixel-level masks, guaranteeing that decisions are interpretative and clinically reliable.
- **State-of-the-art Performance across Multiple Benchmarks**: We achieve state-of-the-art performance across multiple medical VQA benchmarks, such as VQA-RAD, SLAKE and PathVQA. The levels of performance of MedVCoT in fine-grained diagnostic tasks are vastly superior to those of currently available open-source and closed-source models.

2 Related Work

Evolution of Medical Vision-Language Models. Medical VQA originated from contrastive learning paradigms, including BiomedCLIP [Zhang *et al.*, 2023] and PubMedCLIP [Eslami *et al.*, 2023], which provided strong baselines of retrieval and classification [Radford *et al.*, 2021]. The rise of general VLMs (e.g., GPT-4o [Hurst *et al.*, 2024], Gemini 2.0 [Comanici *et al.*, 2025], Qwen2.5-VL [Bai *et al.*, 2025], InternVL3 [Zhu *et al.*, 2025]) further strengthened specialized medical models, such as LLaVA-Med [Li *et al.*, 2023; Liu *et al.*, 2023], BiMediX2 [Mullappilly *et al.*, 2024], HuatuoGPT-Vision [Chen *et al.*, 2024], UniMedVL [Ning *et al.*, 2025], MedDr [He *et al.*, 2024], EchoGPT [Xu *et al.*, 2025b] and Lingshu [Xu *et al.*, 2025c], these powerful backbones were adapted by domain-specific instruction tuning or data reformatting. Nonetheless, they consider medical VQA as an image captioning problem and lack explicit grounding for lesion localization. In contrast, our method introduces a Visual Chain of Thought, where an expert for medical visual reasoning (MedSAM [Ma *et al.*, 2024]) steers the VLM towards the explicit generation of continuous visual tokens while reasoning. The generated tokens can be decoded into segmentation masks, bridging the modality gap between textual reasoning and fine-grained visual perception.

Evolution of Chain-of-Thought Reasoning. The Chain-of-Thought (CoT) paradigm [Wei *et al.*, 2022] has transformed the way that LLMs perform reasoning by embedding intermediate steps in few-shot exemplars. This paradigm has been quickly advanced with progressions including self-consistency decoding, least-to-most decomposition, and Tree of Thoughts (ToT) [Wang *et al.*, 2022; Zhou *et al.*, 2022; Yao *et al.*, 2023], as well as generalization to medical, legal and coding domains [Liu *et al.*, 2024; Servantez *et al.*, 2024; Li *et al.*, 2025]. GLM-4.1V-Thinking [Hong *et al.*, 2025] uses reinforcement learning with discrete textual chain-of-thought

reasoning. Nevertheless, discrete text spaces have difficulty in representing fine-grained visual information for multimodal tasks. To overcome this "modality gap," recent work has turned to latent space reasoning: AURORA [Bigverdi *et al.*, 2025] and Mirage [Yang *et al.*, 2025] introduced perception tokens and machine mental imagery for enhanced visual detail, whereas Chain-of-Visual-Thought (CoVT) [Qin *et al.*, 2025] established an autoregressive framework based on continuous latents for end-to-end visual reasoning. Inspired by these advancements, we propose MedVCoT by introducing this visual chain-of-thought mechanism to the medical domain. Our method bridges visual information and textual diagnostic reasoning, leading to dramatically improved interpretability and accuracy.

3 Method

3.1 Problem Formulation

Let I denote a medical image, Q a medical question, A the ground truth answer, and M ground truth segmentation masks decoded by a medical visual expert (MedSAM). Then our aim is to model the joint generation probability $P(A, T_{vis}|I, Q)$. Here, T_{vis} is a series of continuous latent visual tokens implicitly generated explicitly within a thought horizon (e.g., <think>...</think>). Unlike original VQA, which directly maps $P(A|I, Q)$, MedVCoT forces the T_{vis} to serve as an interpretable bottleneck, these tokens must be decodable into a predicted mask $\hat{M} \approx M_{gt}$ by a visual expert, thus grounding the final textual answer A in verifiable visual reasoning.

3.2 Overall Framework

We propose MedVCoT, a medical vision-language framework that incorporates explicitly visual reasoning in the diagnosis. The architecture consists of three main parts: a fine-tuned Vision-Language Backbone (Qwen2.5-VL-7B), a Latent Visual Projector, and a Medical Visual Expert (MedSAM). The projector is conceived as a lightweight cross-attention module where it leverages a linear layer to project the latent visual tokens of a visual expert into keys and values for attention with a set of learnable queries q, yielding aligned visual embeddings. The overall framework and the main workflow of MedVCoT are depicted in Figure 1.

Given a medical image I and a medical question Q, the backbone firstly processes inputs to trigger a Visual Chain-of-Thought (Visual CoT) mechanism. Instead of just generating the textual answer, the model can autonomously generate a sequence of continuous Latent Visual Tokens T_{vis} enclosed within '' tags. These tokens represent a dense and semantically rich representation of the spatial characteristics of the lesion in the latent space.

Subsequently, T_{vis} is projected by the projector to the prompt embedding space of the frozen MedSAM decoder. MedSAM then decodes these embeddings to a fine-grained predicted segmentation mask $\hat{M}$ with explicitly visual grounding of the model's focus. Lastly, given the input via context (I, Q) as well as the generated visual rationale (T_{vis}), the backbone generates the final textual answer A. This makes the diagnosis not just a statistical hallucination but with a causal connection to verifiable visual evidence $\hat{M}$, and it consolidates perception, reasoning, and generation into a unified framework.

Algorithm 1 Training Process of MedVCoT

Input: Training dataset $\mathcal{D} = \{(I_i, Q_i, A_i)\}_{i=1}^{N}$, VLM backbone $\mathcal{M}_\theta$, Projector $\mathcal{P}_\phi$ (Cross-Attention), Medical Visual Expert (MedSAM) consisting of Encoder $\mathcal{E}_{sam}$ and Decoder $\mathcal{D}_{sam}$.
Parameter: Learnable Query Vectors $\mathbf{q}$ ($N = 8$), Learning rate η, Loss weights λ_{vis}
Output: Optimized parameters $\theta^*, \phi^*, \mathbf{q}^*$

Initialize: Load pre-trained weights for $\mathcal{M}_\theta$ and MedSAM; Initialize $\mathcal{P}_\phi$ and $\mathbf{q}$.
Freeze: Vision Tower of $\mathcal{M}_\theta$, $\mathcal{E}_{sam}$, $\mathcal{D}_{sam}$.
Trainable: θ_{LoRA}, $\mathcal{P}_\phi$, $\mathbf{q}$.

Stage 1: Medical Feature Alignment

for each batch (I, Q, A) in $\mathcal{D}$ **do**
  Construct prompt: $X_{in} = [I, Q, \mathcal{E}_{sam}(I)]$.
  Forward pass: $A_{pred}, H_{vis} = \mathcal{M}_\theta(X_{in})$.
    // H_{vis}: hidden states of $\mathcal{E}_{sam}(I)$
  Visual projection:
  $T_{proj} = \mathcal{P}_\phi(H_{vis}, \mathbf{q})$; $\hat{M} = \mathcal{D}_{sam}(T_{proj})$.
  Get GT masks: $M_{gt} = \text{Filter}(\text{MedSAM}(I))$.
  $\mathcal{L}_{stage1} = \mathcal{L}_{text}(A_{pred}, A) + \lambda_{vis}\mathcal{L}_{vis}(\hat{M}, M_{gt})$.
  Update parameters: $\theta^*, \phi^*, \mathbf{q}^* \leftarrow \text{Optimizer}(\mathcal{L}_{stage1})$.
end for

Stage 2: Visual Reasoning Learning

for each batch (I, Q, A) in $\mathcal{D}$ **do**
  Generate sequence with visual thoughts:
  $A_{pred}, T_{vis} = \mathcal{M}_\theta(I, Q)$
  Project thoughts to visual embeddings:
  $T_{proj} = \mathcal{P}_\phi(T_{vis}, \mathbf{q})$.
  Decode predicted masks: $\hat{M} = \mathcal{D}_{sam}(T_{proj})$.
  Get GT masks: $M_{gt} = \text{Filter}(\text{MedSAM}(I))$.
  $\mathcal{L}_{stage2} = \mathcal{L}_{text}(A_{pred}, A) + \lambda_{vis}\mathcal{L}_{vis}(\hat{M}, M_{gt})$.
  Update parameters: $\theta^*, \phi^*, \mathbf{q}^* \leftarrow \text{Optimizer}(\mathcal{L}_{stage2})$.
end for

Stage 3: Instruction Tuning with Dynamic VCoT

for each batch (I, Q, A) in $\mathcal{D}$ **do**
  Sample Bernoulli indicator: $k \sim \text{Bernoulli}(p)$.
  $\mathcal{L}_{stage3} = \mathcal{L}_{text} + k \cdot \lambda_{vis}\mathcal{L}_{vis}$.
  Update parameters: $\theta^*, \phi^*, \mathbf{q}^* \leftarrow \text{Optimizer}(\mathcal{L}_{stage3})$.
end for
return $\theta^*, \phi^*, \mathbf{q}^*$

3.3 Training Pipeline

In order to effectively endow the model with medical-visual reasoning capability, we propose a progressive three-stage training strategy (shown in Figure 2). During all stages, the MedSAM expert is frozen to keep its pre-trained medical seg-

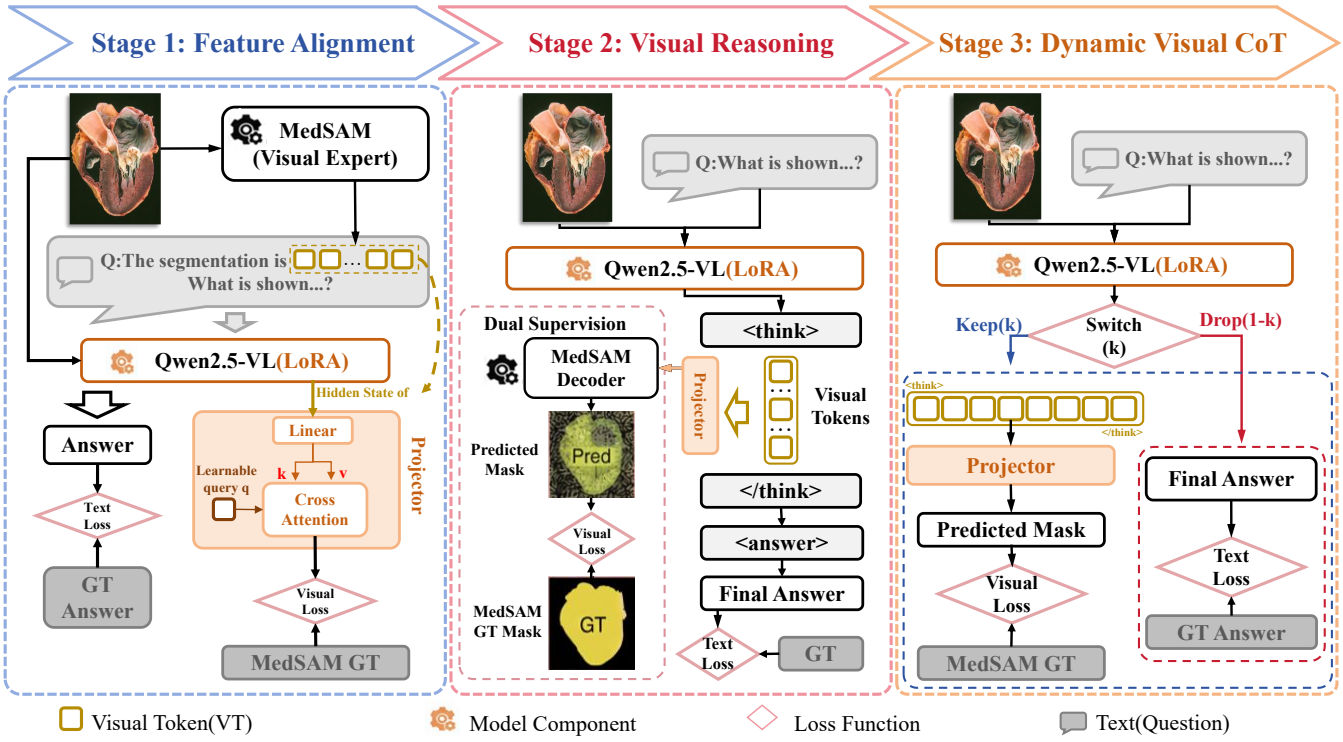

Figure 2: Training Pipeline of MedVCoT.

mentation ability.

Stage 1: Medical Feature Alignment. In this stage, we want the visual encoder's representations to be aligned with the LLM's semantic space. The model processes medical image-text pairs, where the input question accompanies the special tokens (e.g., <sam_pad>). The output hidden states from the LLM are divided into two streams: (1) Textual Stream: The states for textual answer tokens are processed by the language model head to compute the **Text Loss** $\mathcal{L}_{text}$ using a common Cross-Entropy loss. (2) Visual Stream: The states for the special visual tokens are extracted and mapped by the projector to prompt the MedSAM decoder, and then segmentation masks are generated. Such predictions are guided by the **Visual Loss** $\mathcal{L}_{vis}$ with respect to ground-truth masks. This dual-task objective effectively provides a way to align visual features and diagnostic semantics in the beginning.

The total loss for Stage 1 is:

$$\mathcal{L}_{Stage1} = \mathcal{L}_{text} + \lambda_{vis}\mathcal{L}_{vis}$$

Here, $\mathcal{L}text$ is the standard Cross-Entropy loss for next-token prediction, and λ_{vis} is a hyperparameter balancing the textual and visual objectives. Simultaneously, the Visual Loss $\mathcal{L}_{vis}$ supervises the quality of the generated visual tokens by comparing the decoded masks $\hat{M}$ with the ground-truth masks M_{gt}(automatically distilled from MedSAM with stability-based quality filtering):

$$\mathcal{L}_{vis} = \mathcal{L}_{Dice}(\hat{M}, M_{gt}) + \lambda_{focal}\mathcal{L}_{Focal}(\hat{M}, M_{gt})$$

Here, $\mathcal{L}_{Dice}$ guarantees global shape overlap, while $\mathcal{L}_{Focal}$ forces the model to focus on hard-to-classify boundary pixels.

Stage 2: Visual Reasoning Learning. In this stage, we integrate visual representation learning and diagnostic reasoning into one training process. We guide the model to autonomously generate explicit visual evidence before generating the ultimate diagnosis. In particular, the model is required to generate a structured output sequence: it first generates continuous latent visual tokens in order to capture pixel-level lesion traits and then uses these variables as rationale-grounded evidence to generate the textual answer. We freeze the vision tower and update our projector and query vectors, as well as LoRA parameters. Meanwhile, we employ a joint optimization strategy in order to balance between visual quality and diagnostic accuracy. The overall loss for Stage 2 is given by:

$$\mathcal{L}_{Stage2} = \mathcal{L}_{text} + \lambda_{vis}\mathcal{L}_{vis}$$

where λ_{vis} balances the two objectives (set to 1.0 by default). The Text Loss $\mathcal{L}_{text}$ (Cross-Entropy) ensures the model follows the instruction to generate the correct reasoning path and final answer. This hybrid objective enforces a strict causal dependency: the model must "see" the lesion correctly (via $\mathcal{L}_{vis}$) to answer the question correctly (via $\mathcal{L}_{text}$).

Stage 3: Instruction Tuning using Dynamic Visual CoT. In this stage, to facilitate the model for application in complex clinical scenarios and improve its robustness, we conduct instruction tuning for MedVCoT on fine-grained medical clinical VQA datasets. We introduce a dynamic visual CoT ap-

Model	VQA-RAD		VQA-Med	SLAKE		PathVQA		Overall
	All	Open	All	All	Open	All	Open	
Closed-Source								
GPT-4o [Hurst *et al.*, 2024]	54.5	46.9	60.0	68.3	65.1	31.3	16.7	41.8
Gemini2.0-flash [Comanici *et al.*, 2025]	58.3	**58.1**	59.6	70.1	66.1	41.8	26.5	49.5
Open-Source(<10B)								
Bimedix2-8B [Mullappilly *et al.*, 2024]	53.4	36.9	49.6	54.3	54.4	28.6	9.0	36.3
Llava-Med-v1.5-7B [Li *et al.*, 2023]	47.2	31.8	38.0	48.9	42.4	34.3	6.8	37.6
Qwen2.5-VL-7B [Bai *et al.*, 2025]	61.9	46.4	48.4	65.3	59.5	34.8	9.5	43.3
HuatuoGPT-Vision-7B [Chen *et al.*, 2024]	64.3	49.7	58.4	64.5	59.1	36.3	12.3	44.8
InternVL3-8B [Zhu *et al.*, 2025]	51.0	35.8	58.8	70.3	64.9	39.4	12.8	47.5
GLM-4.1V-9B-Thinking [Hong *et al.*, 2025]	60.1	43.0	55.6	67.2	64.7	46.8	31.9	52.2
Lingshu-7B [Xu *et al.*, 2025c]	67.2	55.9	57.4	<u>80.6</u>	<u>78.9</u>	<u>58.4</u>	<u>29.6</u>	<u>63.5</u>
Open-Source(>10B)								
HuatuoGPT-Vision-34B [Chen *et al.*, 2024]	65.6	52.5	**62.8**	67.1	62.4	36.0	12.4	45.4
InternVL3-14B [Zhu *et al.*, 2025]	45.5	25.7	59.0	70.5	66.5	41.0	13.1	48.4
Qwen2.5-VL-32B [Bai *et al.*, 2025]	49.2	46.4	57.8	75.6	70.2	39.3	14.0	48.5
MedDr-40B [He *et al.*, 2024]	59.2	39.1	47.4	65.8	58.2	45.8	15.9	50.8
UniMedVL-14B [Ning *et al.*, 2025]	<u>68.5</u>	48.0	59.2	77.9	70.8	51.9	26.7	58.7
Ours								
MedVCoT-7B (Ours)	**69.2** **+1.0%**	<u>57.5</u> **-1.0%**	<u>61.4</u> **-2.2%**	**82.8** **+2.8%**	**83.5** **+5.8%**	**64.0** **+9.6%**	**35.9** **+21.1%**	**68.1** **+7.2%**

Table 1: Comparison of different models on medical VQA datasets. 'All' denotes overall accuracy, 'Open' denotes accuracy on open-ended questions. Best results are **bold**, second best are <u>underlined</u>.

proach mechanism with a Bernoulli-driven strategy. This procedure stochastically decides whether it should generate the full sequence of visual reasoning tokens or if it will opt out of that for straight text inference. This promotes the model to be flexible about whether or not explicit visual grounding is desired, while preserving the opportunity for direct diagnosis if required. The optimization target for this stage incorporates a dynamic indicator function $k \sim \text{Bernoulli}(p)$, where p represents the probability of activating the visual reasoning path.

$$\mathcal{L}_{Stage3} = \mathcal{L}_{text} + k \cdot \lambda_{vis} \mathcal{L}_{vis}$$

When $k = 1$ (Keep): The model generates visual tokens, and the loss includes both text supervision $\mathcal{L}_{text}$ and visual reconstruction supervision $\mathcal{L}_{vis}$. When $k = 0$ (Drop): The visual generation path is deactivated, and the model is optimized solely via $\mathcal{L}_{text}$, simulating scenarios where visual evidence is implicit or reasoning is purely linguistic.

4 Experiments

4.1 Datasets and Evaluation Metrics

We evaluate our method on four diverse medical VQA benchmarks: **VQA-RAD** [Lau *et al.*, 2018], a manually curated radiology dataset containing 315 images and 3,515 QA pairs verified by clinicians; **SLAKE** [Liu *et al.*, 2021], a bilingual dataset with 642 annotated images (CT/MRI/X-ray) and 14k QA pairs, incorporating rich semantic labels; **PathVQA** [He *et al.*, 2020], the first large-scale pathology VQA dataset with 4,998 images and over 32k open-ended QA pairs derived from medical textbooks; and **VQA-Med 2019** [Ben Abacha *et al.*, 2019], a radiology-focused challenge dataset with 4,200 images and 15k QA pairs covering modality, plane, organ, and abnormality categories. For all datasets, we adopt a standard random split, using 80% of the data for training and 20% for testing. All experiments were conducted on a server equipped with two NVIDIA A800 GPUs.

Primary metrics include overall ('All') and open-ended ('Open') accuracy. To ensure robust evaluation against medical terminology variations, we utilize GPT-4o as a semantic judge, explicitly instructing it to *"determine if the prediction is semantically consistent with the ground truth, prioritizing core meaning over phrasing."* This protocol effectively handles clinical synonyms and ignores irrelevant surface-level discrepancies.

4.2 Main Results

We benchmarked MedVCoT against 14 leading models, including closed-source (GPT-4o, Gemini 2.0) and open-source (Qwen2.5-VL, LLaVA-Med, Lingshu, etc.) baselines, across four medical VQA datasets. For a fair comparison, all baseline models were evaluated using their official pre-trained weights directly on the test sets under a unified evaluation protocol.

As shown in Table 1, MedVCoT achieves superior overall performance with an average accuracy of 68.1%, outperform-

ing all competitors, including the previous SOTA open-source model Lingshu-7B (+7.2%) and the closed-source Gemini 2.0 (+37.5% relative improvement). This advantage is particularly pronounced in tasks requiring fine-grained visual perception. On SLAKE, which features rich semantic annotations, our model achieves an Open accuracy of 83.5%, significantly surpassing Lingshu (78.9%) and UniMedVL (70.8%). Notably, on the challenging PathVQA dataset, our Open accuracy reaches 35.9%, a substantial 21.1% improvement over the runner-up Lingshu (29.6%). These results strongly demonstrate that incorporating an explicit Visual Chain-of-Thought effectively bridges the gap in lesion localization and fine-grained feature recognition inherent in traditional VLMs. Although slightly outperformed by HuatuoGPT-Vision-34B on VQA-Med (-2.2%), this marginal gap is primarily attributable to the significant disparity in model parameter scale (7B vs. 34B), yet our model maintains exceptional competitiveness in open-ended reasoning tasks even with a much lighter architecture.

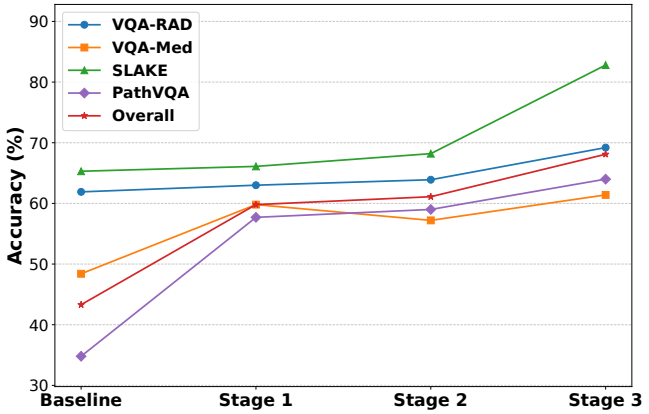

Figure 3: Ablation study on different training stages across four datasets.

4.3 Ablation Study

Based on the ablation study visualized in Figure 3, we analyze the cumulative impact of our three-stage training strategy compared to the baseline [Bai *et al.*, 2025; Qin *et al.*, 2025]. The baseline model, lacking domain-specific adaptation, achieves a modest overall accuracy of 43.3%.

Stage 1 (Feature Alignment) provides the most significant initial boost, propelling overall accuracy to 59.8% (+16.5%). This sharp rise is particularly evident in PathVQA (+22.9%) and VQA-Med (+11.4%), indicating that establishing a foundational mapping between the VLM latent space and medical visual features is crucial for basic domain competency.

Stage 2 (Visual Reasoning Learning) introduces the Visual CoT mechanism, yielding steady gains (overall +1.3% to 61.1%). While improvements on VQA-RAD and PathVQA confirm the benefit of reasoning capabilities, a slight dip in VQA-Med suggests that learning to reason without task-specific tuning may introduce complexity that temporarily affects performance on simpler tasks.

Stage 3 (Medical Tuning) delivers the final qualitative leap, surging overall accuracy to 68.1% (+7.0%). The most dramatic improvement is observed on SLAKE (+14.6%), where instruction tuning and robust training strategies (e.g., random token dropping) fully unlock the model's potential for complex, semantically rich queries. This trajectory (43.3% $\rightarrow$ 59.8% $\rightarrow$ 61.1% $\rightarrow$ 68.1%) confirms that while feature alignment lays the groundwork, the integration of explicit visual reasoning and domain-specific tuning is essential for achieving state-of-the-art performance.

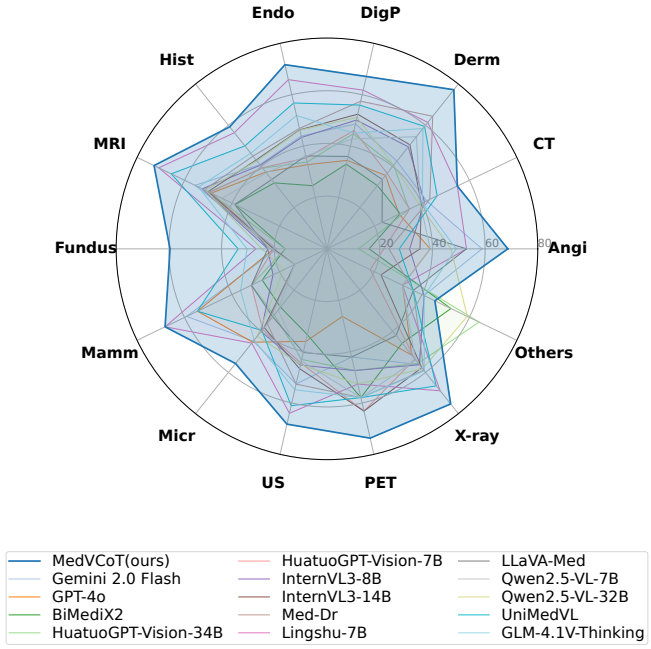

Figure 4: The radar chart illustrates the accuracy comparison of MedVCoT and other SOTA models across diverse medical modalities. Abbreviations: DigP (Digital Photography), Hist (Histopathology), Mamm (Mammography), Angi (Angiography), Endo (Endoscopy), Derm (Dermoscopy), Micr (Microscopy), US (Ultrasound).

4.4 Fine-grained Capability Analysis

To rigorously assess robustness and generalization boundaries, we extend beyond aggregate metrics to conduct a fine-grained capability analysis based on imaging modalities. We categorize test samples into 14 distinct sub-categories , spanning macroscopic imaging (e.g., Dermoscopy, Endoscopy), microscopic pathology (e.g., Histopathology), diverse radiological modalities (e.g., CT, MRI, X-ray), and a miscellaneous 'Others' class. This stratified evaluation comprehensively profiles the model's consistency across varying imaging mechanisms and anatomical structures, elucidating its cross-modal strengths.

As illustrated in Figure 4, MedVCoT (dark blue line) demonstrates superior comprehensive performance, with its envelope area significantly encompassing all baselines. Specifically, our model maintains a decisive lead in common modalities such as MRI, CT, and Fundus, while exhibiting exceptional stability in challenging tasks like Histopathology.

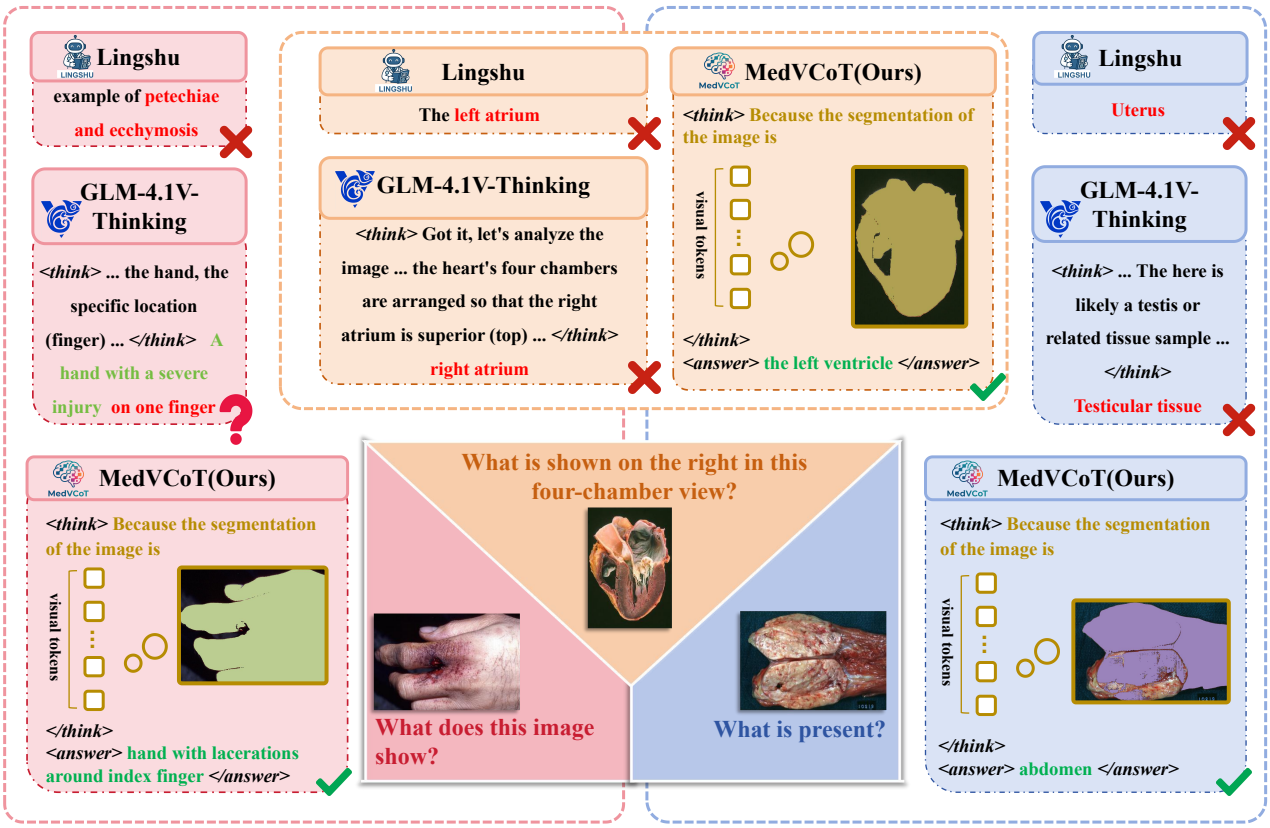

Figure 5: Qualitative comparison between MedVCoT and baseline models (Lingshu-7B, GLM-4.1V-9B-Thinking). MedVCoT generates latent visual tokens that can be decoded into visible masks, grounding the diagnosis in verifiable pixel-level evidence and significantly enhancing interpretability.

Although a slight performance dip is observed in the 'Others' category, this is attributable to its extreme data scarcity (accounting for only 0.22% of the dataset), which has negligible impact on overall clinical utility. These results strongly suggest that the explicit Visual Chain-of-Thought effectively enhances perceptual sensitivity to diverse features, providing reliable decision support across multi-disciplinary diagnostics.

4.5 Case Studies

To qualitatively assess the reasoning logic in authentic clinical scenarios, we select three representative and challenging cases covering anatomical localization, injury differentiation, and specimen identification, comparing MedVCoT against two strong baselines. As illustrated in Figure 5, baseline models, despite incorporating textual CoT capabilities, still lack explicit visual grounding mechanisms and tend to rely on surface-level feature matching or hallucinations. For instance, in the first case, both baselines failed to correctly distinguish the anatomical orientation of the four-chamber view, consistently misidentifying the structure as an atrium rather than the ventricle. In the second case, they misclassified obvious lacerations as petechiae or provided vague descriptions lacking diagnostic precision. Furthermore, in the specimen identification case, both baselines hallucinated reproductive organs (uterus or testicular tissue) instead of correctly recognizing the abdominal tissue.

In contrast, MedVCoT, by activating a Visual Chain-of-Thought (Visual CoT) within the latent space, first generates precisely aligned anatomical segmentation tokens. This process acts as a visual introspection prior to answering. This mechanism enables the model to explicitly localize key evidence and structurally connect it with global semantics, resulting in correct predictions across all cases. These qualitative results robustly validate that incorporating visual reasoning steps in the latent space not only improves diagnostic accuracy but also significantly enhances the interpretability and transparency of the model's decision-making process.

5 Conclusion

This study successfully develops and validates MedVCoT, a medical VQA framework that bridges modality gap via a Visual Chain-of-Thought. By generating continuous latent visual tokens, MedVCoT enhances both diagnostic accuracy and clinical interpretability through verifiable segmentation masks. Our model demonstrates state-of-the-art performance across four authoritative benchmarks. Future work will explore scaling MedVCoT to larger-parameter models to verify its universality and integrate diverse visual experts to optimize the Visual CoT.

Ethical Statement

There are no ethical issues.

Acknowledgments

This work is supported by the Science and Technology Project of Liaoning Province (2023JH2/101700363, 2024JH2/102600027), in part by the National Natural Science Foundation of China under Grant 62072073, 62106034, in part by the Fundamental Research Funds for the Central Universities under Grant DUT24ZD124, in part by the Dalian Innovation Fund 2021JJ12GX016, and the Science and Technology Project of Dalian City (2024JJ12GX025, 2023JJ12SN029 and 2023JJ11CG005).

References

[Bai *et al.*, 2025] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[Ben Abacha *et al.*, 2019] Asma Ben Abacha, Sadid A Hasan, Vivek V Datla, Dina Demner-Fushman, and Henning Müller. 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.

[Bigverdi *et al.*, 2025] Mahtab Bigverdi, Zelun Luo, Cheng-Yu Hsieh, Ethan Shen, Dongping Chen, Linda G Shapiro, and Ranjay Krishna. Perception tokens enhance visual reasoning in multimodal language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3836–3845, 2025.

[Chen *et al.*, 2024] Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, et al. Towards injecting medical visual knowledge into multimodal llms at scale. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 7346–7370, 2024.

[Comanici *et al.*, 2025] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[Eslami *et al.*, 2023] Sedigheh Eslami, Christoph Meinel, and Gerard De Melo. Pubmedclip: How much does clip benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1181–1193, 2023.

[He *et al.*, 2020] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

[He *et al.*, 2024] Sunan He, Yuxiang Nie, Zhixuan Chen, Zhiyuan Cai, Hongmei Wang, Shu Yang, and Hao Chen. Meddr: Diagnosis-guided bootstrapping for large-scale medical vision-language learning. *CoRR*, 2024.

[Hong *et al.*, 2025] Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

[Hurst *et al.*, 2024] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[Lau *et al.*, 2018] 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.

[Li *et al.*, 2023] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023.

[Li *et al.*, 2025] Jia Li, Ge Li, Yongmin Li, and Zhi Jin. Structured chain-of-thought prompting for code generation. *ACM Transactions on Software Engineering and Methodology*, 34(2):1–23, 2025.

[Liu *et al.*, 2021] 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.

[Liu *et al.*, 2023] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[Liu *et al.*, 2024] Jiaxiang Liu, Yuan Wang, Jiawei Du, Joey Tianyi Zhou, and Zuozhu Liu. Medcot: Medical chain of thought via hierarchical expert. *arXiv preprint arXiv:2412.13736*, 2024.

[Ma *et al.*, 2024] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

[Mullappilly *et al.*, 2024] Sahal Shaji Mullappilly, Mohammed Irfan Kurpath, Sara Pieri, Saeed Yahya Alseiari, Shanavas Cholakkal, Khaled Aldahmani, Fahad Khan, Rao Anwer, Salman Khan, Timothy Baldwin, et al. Bimedix2: Bio-medical expert lmm for diverse medical modalities. *arXiv preprint arXiv:2412.07769*, 2024.

[Ning *et al.*, 2025] Junzhi Ning, Wei Li, Cheng Tang, Jiashi Lin, Chenglong Ma, Chaoyang Zhang, Jiyao Liu, Ying

Chen, Shujian Gao, Lihao Liu, et al. Unimedvl: Unifying medical multimodal understanding and generation through observation-knowledge-analysis. *arXiv preprint arXiv:2510.15710*, 2025.

[Qin *et al.*, 2025] Yiming Qin, Bomin Wei, Jiaxin Ge, Konstantinos Kallidromitis, Stephanie Fu, Trevor Darrell, and Xudong Wang. Chain-of-visual-thought: Teaching vlms to see and think better with continuous visual tokens. *arXiv preprint arXiv:2511.19418*, 2025.

[Radford *et al.*, 2021] 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.

[Servantez *et al.*, 2024] Sergio Servantez, Joe Barrow, Kristian Hammond, and Rajiv Jain. Chain of logic: Rule-based reasoning with large language models. *arXiv preprint arXiv:2402.10400*, 2024.

[Wang *et al.*, 2022] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

[Wei *et al.*, 2022] 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.

[Xu *et al.*, 2025a] Bo Xu, Jinshi Yu, Hao Li, Quanhao Zhu, Zihan Bai, and Chuansen Yuan. Geneagent: An interpretable oncogene prediction and evidence-based question-answering framework. In *2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 5962–5969. IEEE, 2025.

[Xu *et al.*, 2025b] Bo Xu, Quanhao Zhu, Qingchen Zhang, Mengmeng Wang, Liang Zhao, Hongfei Lin, Jing Ren, and Feng Xia. Echogpt: An interactive cardiac function assessment model for echocardiogram videos. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 7858–7866, 2025.

[Xu *et al.*, 2025c] Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*, 2025.

[Yang *et al.*, 2025] Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. *arXiv preprint arXiv:2506.17218*, 2025.

[Yao *et al.*, 2023] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

[Zhang *et al.*, 2023] 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.

[Zhou *et al.*, 2022] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2022.

[Zhu *et al.*, 2025] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.