When Models Learn to Ask Why: Adaptive Causal Reasoning for Trustworthy Medical Vision–Language Models

Jianxin Lin[1,*], Chunzheng Zhu[2,*], Peter J. Kneuertz[1], Yunfei Bai[3,‡], Yuan Xue[1,†]
[1]The Ohio State University [2]Hunan University [3]Amazon
xue.643@osu.edu

1. Additional Implementation Details

1.1. Detailed Experimental Configuration

MedCausalX employs Qwen2.5-VL-32B [1] as the base architecture with parameter-efficient LoRA adaptation targeting query, key, and value projection layers. Data augmentation preserves diagnostic integrity through random rotation (± 15), horizontal flipping (probability 0.5), brightness/contrast adjustment ($\pm 20\%$), and Gaussian noise injection ($\sigma = 0.03$). The framework augments vocabulary with reflective tokens $\langle \texttt{CAUSAL} \rangle$ and $\langle \texttt{VERIFY} \rangle$, enabling all training samples to follow unified structure $(I, q, \langle \texttt{CAUSAL} \rangle, R^{\text{biased}}, \langle \texttt{VERIFY} \rangle, R^{\text{gold}}, y)$ for learning adaptive verification through reasoning trajectories.

Training proceeds through two-stage optimization following preliminary supervised fine-tuning. Stage I (Causal SFT) maximizes joint likelihood over anatomical localization, causal chains, and diagnostic predictions for 3 epochs using AdamW with learning rate $\eta = 1 \times 10^{-6}$, weight decay 5×10^{-5}, gradient clipping threshold 1.0, and 500-step linear warmup followed by cosine annealing. Stage II integrates dual-policy refinement: off-policy DPO [12] over 2 epochs with KL coefficient $\beta = 0.1$ leverages error trajectories where semantic similarity threshold $\tau = 0.7$ identifies failure points t_{fail} via $\mathcal{S}(Y^t_{\text{err}}, Y^t_{\text{gt}})$ comparison, constructing preference pairs contrasting erroneous continuations against correct trajectories; on-policy GRPO [13] over 2000 steps samples $G = 8$ trajectories per input, computing group-relative advantages $A^{(g)} = (R^{(g)} - \bar{R})/(\sigma_R + \epsilon)$ with stability constant $\epsilon = 1 \times 10^{-8}$. Distributed training across 6 NVIDIA A100 (40GB) GPUs with gradient accumulation (4 steps) and mixed precision (bfloat16) converges within approximately 56 hours per dataset using 5-fold cross-validation with stratified 8:1:1 splits. Table 1 details the complete configuration.

1.2. Benchmark Dataset Statistics

We evaluate MedCausalX on established medical imaging benchmarks spanning multiple modalities and clinical domains. *MIMIC-CXR* [5] provides 377,110 chest radiographs from 227,835 imaging studies with free-text reports and 14 pathology labels for report generation tasks. *SLAKE* [9] contains 642 multi-modal images with 14,028 question-answer pairs across anatomy, modality, and pathology categories. *VQA-RAD* [7] includes 315 radiology images with 3,515 QA pairs emphasizing clinical reasoning. *PathVQA* [3] comprises 32,799 pathology images with 6,719 QA pairs for histopathological interpretation. *SA-Med2D-20M* [15] offers 4.6 million 2D medical images with 19.7 million segmentation masks across 10 modalities and 31 organs, enabling region-centric tasks including structure identification and lesion localization with bounding box annotations. All experiments employ 5-fold cross-validation using stratified 8:1:1 train-validation-test splits for statistical reliability.

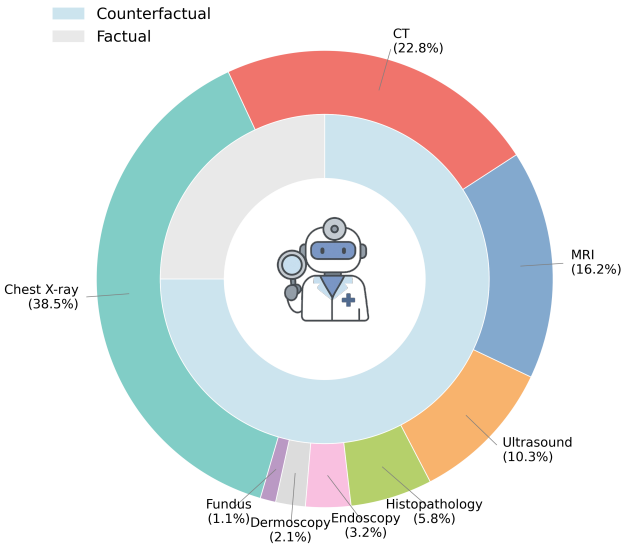


Figure 1. Modality distribution and factual-counterfactual composition of CRMed dataset. The dataset spans multiple imaging modalities with constructed counterfactual variants through controlled interventions, enabling robust causal reasoning evaluation.

CRMed Dataset Construction. The CRMed dataset aug-

* Equal contribution. † Corresponding author. ‡ Work conducted in an individual capacity and does not reflect the views of Amazon.

Table 1. Complete hyperparameter configuration for MedCausalX training and inference.

Hyperparameter	Value
Base Model	Qwen2.5-VL-32B
LoRA Rank (r)	64
LoRA Alpha (α)	128
LoRA Dropout	0.05
Target Modules	q_proj, k_proj, v_proj
Reflective Tokens	⟨CAUSAL⟩, ⟨VERIFY⟩
Nucleus Sampling (p)	0.9
Temperature (T)	0.7
Batch Size	64
Max Tokens ($T_{\max}$)	2048
Causal SFT (Stage I)	
Epochs	3
Optimizer	AdamW
Learning Rate (η)	1×10^{-6}
Weight Decay	5×10^{-5}
Gradient Clipping	1.0
Warmup Steps	500
LR Schedule	Linear + Cosine
Dual-Policy Optimization (Stage II)	
Off-policy training:	
Epochs	2
KL Coefficient (β)	0.1
Error Threshold (τ)	0.7
On-policy training:	
Training Steps	2000
Group Size (G)	8
Stability Constant (ϵ)	1×10^{-8}
Infrastructure	
GPUs	6× A100 (40GB)
Gradient Accumulation	4 steps
Mixed Precision	bfloat16
Training Time	∼56 hours/dataset
Cross-validation	5-fold
Data Split	8:1:1

ments existing benchmarks with two systematic annotation layers. *First*, we select images from SA-Med2D-20M and MIMIC-CXR based on multi-criteria quality assessment: (i) sufficient image resolution ($\geq$512×512 pixels) and contrast quality (SNR>15dB), (ii) presence of diagnostically relevant anatomical structures confirmed through clinical relevance scoring ($\geq$4/5 by three board-certified radiologists). Selected images include bounding box annotations in format $(x_{\min}, y_{\min}, x_{\max}, y_{\max})$ covering target anatomical structures. *Second*, counterfactual variants generate $\mathcal{D}_{\text{shortcut}}$ via spatial perturbations (random shifts $\pm$50 pixels, scale jittering $\times[0.8, 1.2]$) disrupting $A \to P$ dependencies, and $\mathcal{D}_{\text{partial}}$ through pathology misalignment breaking $P \to Y$ causal flow while preserving image content. Intervention quality is validated by ensuring IoU>0.7 between perturbed and original bounding boxes to maintain anatomy. The resulting corpus contains 89,342 images with 267,128 causal samples, averaging 4.2 reasoning steps per chain (SD=1.3). Modality distribution spans chest X-ray (38.5%), CT (22.8%), MRI (16.2%), ultrasound (10.3%), histopathology (5.8%), endoscopy (3.2%), dermoscopy (2.1%), and fundus (1.1%), with counterfactual plausibility scores >0.85 via GPT-4V evaluation.

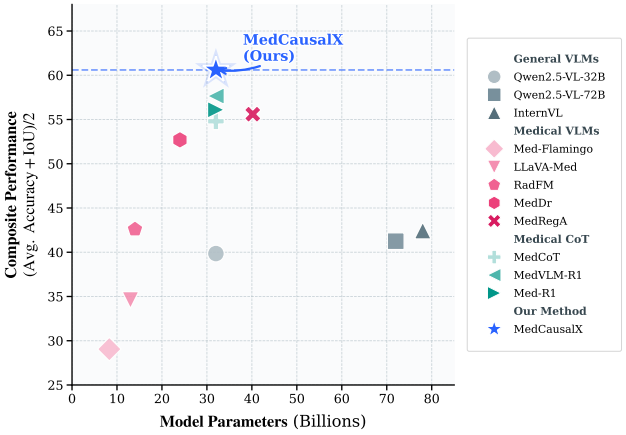


Figure 2. Comparison of model efficiency and performance across medical VLMs. MedCausalX achieves the highest composite score at 32B parameters, demonstrating superior causal reasoning capabilities while maintaining computational efficiency.

2. Additional Experimental Results

2.1. Causal Reasoning Cases

MedCausalX demonstrates robust causal reasoning across multiple public medical VQA datasets. In Figure 5 (Case 1), the model successfully performs multi-object anatomical quantification, accurately identifying bilateral pulmonary structures via causal factorization. Figure 6 (Case 2) illustrates its capability in anatomical landmark detection, where the presence of the gastric cardia is correctly confirmed through systematic spatial reasoning. In Figure 7 (Case 3), MedCausalX shows precise spatial localization by identifying an endocardial lesion adjacent to the lower intraventricular septum, leveraging structured anatomical factorization ($A \to P \to Y$). Finally, Figure 8 (Case 4) demonstrates binary anatomical detection, with the model reliably confirming uterine presence through causal verification. Together, these examples highlight that MedCausalX grounds its diagnostic predictions in genuine anatomical evidence, rather than relying on superficial pattern correlations.

Table 2. Model scaling analysis across region-centric medical reasoning tasks. We evaluate MedCausalX variants with 3B, 14B, and 32B parameters to assess the impact of model capacity on causal reasoning performance. Best results are highlighted in **bold**, second-best are underlined. All models trained with identical data and optimization strategies.

Model Scale	Region-to-Text Identification										Text-to-Region Detection							
	Structure Identification					Lesion Identification					Single-Region Detection				Multi-Region Detection			
	BLEU-1	F1	Recall	BertScore	Acc	BLEU-1	F1	Recall	BertScore	Acc	Obj-F1	Reg-F1	Align-F1	IoU	Obj-F1	Reg-F1	Align-F1	IoU
MedCausalX-3B	53.42	54.68	55.20	52.15	48.20	38.29	39.12	40.35	48.65	35.12	68.45	21.47	18.23	24.49	55.18	14.23	12.47	19.82
MedCausalX-14B	73.49	74.23	75.80	54.20	55.30	56.43	57.80	59.25	70.83	52.80	79.82	34.62	28.15	33.75	65.75	19.56	17.85	26.18
MedCausalX-32B	**79.82**	**80.15**	**81.25**	**88.45**	**70.15**	**63.27**	**64.35**	**65.80**	**81.90**	**61.35**	**82.45**	**46.82**	**44.92**	**44.35**	**72.38**	**29.51**	**26.82**	**35.76**

Table 3. Model scale and architecture comparison across medical VQA benchmarks. Light blue background indicates selected configuration.

Model	VQA-RAD	SLAKE	PathVQA	PMC-VQA	Average
Vision-Language Models (Zero-shot)					
LLaVA-1.5-13B [10]	53.2	58.5	52.8	56.3	55.2
InternVL-2-40B [2]	59.8	65.2	61.5	67.1	63.4
MedRegA-40B [14]	76.9	84.1	68.5	79.5	77.3
Medical CoT Methods (Qwen2.5-VL-32B Backbone)					
Med-R1 [6]	76.9	85.4	69.5	81.9	78.4
MedVLM-R1 [11]	77.3	85.7	70.8	82.4	79.1
MedCausalX Scale Variants (Qwen2.5-VL Backbone)					
MedCausalX-3B	67.8	75.9	62.5	68.2	68.6
MedCausalX-7B	73.5	81.8	67.3	74.9	74.4
MedCausalX-32B	**79.8**	**87.2**	**73.2**	**84.6**	**81.2**

Table 4. Ablation study on training strategy configuration. Light blue background indicates selected configuration, bold values denote best performance in each row.

Config.	Value	VQA-RAD	SLAKE	PathVQA	PMC-VQA	Average
(a) Training Stage Order						
Seq.	GRPO→DPO	73.5	81.4	68.2	78.9	75.5
	DPO→GRPO	**79.8**	**87.2**	**73.2**	**84.6**	**81.2**
(b) GRPO Group Size G						
G	4	76.8	84.3	70.5	81.7	78.3
	6	78.5	86.1	72.1	83.5	80.1
	8	**79.8**	87.2	**73.2**	**84.6**	**81.2**
	12	78.9	**88.5**	72.6	83.9	80.9
	16	77.4	85.2	71.3	82.4	79.1

2.2. Prompt Engineering and Reflective Tokens

MedCausalX utilizes a structured prompt template (Figure 4) to guide causal reasoning through explicit task decomposition. The system prompt directs the model to systematically examine medical images, compare findings with initial predictions, and link observations to the provided concept vocabulary using concrete visual evidence. This hierarchical framework ensures semantic consistency with medical ontologies while producing declarative sentences suitable for causal verification. Reflective tokens `<causal>` and `<verify>` facilitate dynamic reasoning mode transitions, allowing preliminary anatomical analysis to trigger subsequent verification and correction phases whenever inconsistencies are detected.

2.3. Efficiency-Performance Analysis

Figure 2 compares efficiency and performance of state-of-the-art medical vision-language models using a composite score averaging diagnostic accuracy and spatial grounding IoU. General-purpose VLMs improve modestly with scale but remain below medical-specific models. Domain-adapted medical VLMs achieve higher performance, and chain-of-thought enhanced models further benefit from structured reasoning. MedCausalX sets a new state-of-the-art, outperforming comparable 32B models and even larger specialized systems while using fewer parameters. These results highlight that explicit causal reasoning and structured factorization provide substantial gains beyond raw parameter scaling, establishing MedCausalX as the most efficient and effective solution in its class.

2.4. MedCausalX Scaling Analysis

Table 2 examines how model capacity influences causal reasoning. The 3B variant captures basic anatomical structures but shows clear limitations in lesion-level reasoning, achieving only 53.42% BLEU-1 for structure identification. Scaling to 14B parameters improves performance by 20 approximately points, indicating better causal decomposition and understanding of medical concepts. The 32B variant achieves the best overall results, with BLEU-1 of 79.82% and BertScore of 88.45%, demonstrating strong semantic coherence, precise spatial grounding, and consistent reasoning quality. These trends confirm that larger models enable deeper causal verification and stronger spatial–linguistic alignment, supporting MedCausalX's scalable architecture for complex medical reasoning.

2.5. Model Architecture Effect

Table 3 highlights how architectural design and model capacity affect VQA reasoning. MedCausalX-32B achieves 81.2% average accuracy, outperforming CoT baselines Med-R1 and MedVLM-R1 by 2–3 points, demonstrating that its causal factorization and error-attributed reinforcement learning architecture enables structured reasoning beyond standard chain-of-thought. Performance scales from 68.6% at 3B to 81.2% at 32B, showing that counterfactual reasoning benefits from increased capacity. General VLMs

Table 5. Ablation of reward components and error localization threshold. Light blue background indicates selected configuration.

Config.	Value	VQA-RAD	SLAKE	PathVQA	PMC-VQA	Average
(a) Reward Weight Configuration $\lambda_{acc}/\lambda_{causal}$						
λ	1.0/0.0	71.5	78.9	65.3	76.2	73.0
	0.7/0.3	76.8	84.1	70.5	81.7	78.3
	0.5/0.5	**79.8**	**87.2**	**73.2**	**84.6**	**81.2**
	0.3/0.7	78.2	85.7	71.8	83.1	79.7
	0.0/1.0	67.3	74.5	61.2	71.8	68.7
(b) Error Localization Threshold τ						
τ	0.3	70.8	77.5	64.1	75.3	71.9
	0.5	75.3	82.9	69.4	80.5	77.0
	0.7	**79.8**	**87.2**	**73.2**	**84.6**	**81.2**
	0.8	77.1	84.6	70.8	82.3	78.7
	0.9	72.4	79.8	66.5	77.1	73.9

Table 6. Out-of-domain evaluation on specialized medical VQA datasets. Models trained on core benchmarks and evaluated zero-shot on unseen domains.

Method	OVQA	EndoVis-VQA	Skin-VQA	Average
GPT-4o [4]	43.5	39.8	46.1	43.1
LLaVA-Med [8]	47.2	44.3	49.5	47.0
InternVL-2 [2]	49.8	46.7	51.2	49.2
Med-R1 [6]	44.8	40.5	46.3	43.9
MedVLM-R1 [11]	50.3	47.1	49.8	49.1
MedCausalX (Ours)	**58.4**	**54.2**	**59.7**	**57.4**

LLaVA-1.5 (55.2%) and InternVL-2 (63.4%) lag behind, emphasizing the importance of specialized architecture for anatomically grounded diagnostic tasks.

2.6. Training Strategy Ablation

Table 4 validates our dual-stage reinforcement learning design. The DPO→GRPO sequence achieves 81.2% average accuracy, outperforming the reversed order by 5.7 points. This indicates that applying DPO first allows off-policy learning from error trajectories, effectively establishing robust reasoning foundations, while subsequent GRPO exploration refines causal intervention strategies without premature convergence to superficial patterns. Regarding group size configuration, G=8 provides an optimal balance: smaller groups introduce excessive gradient noise that destabilizes causal learning, whereas larger configurations, such as G=12 or 16, exhibit diminishing returns because conservative advantage normalization suppresses the discovery of novel interventions, which is essential for adaptive and flexible reasoning.

2.7. Reward Component Configuration

Table 5 highlights the interdependence of reward components in causal reasoning optimization. Balanced weighting achieves the best accuracy, outperforming accuracy-only and causal-only setups, showing that multi-objective optimization prevents overfitting to diagnostic shortcuts while preserving reasoning quality. For error localization threshold τ, our study shows that 0.7 provides the best trade-off between sensitivity and specificity: lower thresholds generate excessive false positives, while higher thresholds miss critical reasoning errors, confirming that precise error attribution is essential for effective causal training.

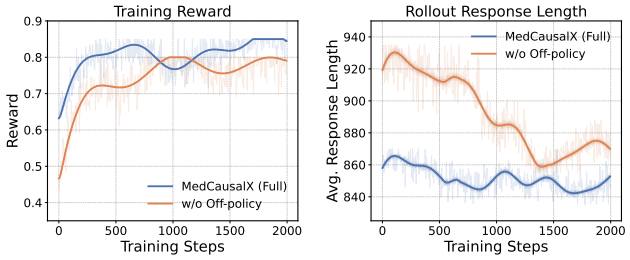


Figure 3. Training dynamics of on-policy GRPO optimization. MedCausalX (Full) with DPO pretraining achieves higher rewards and maintains stable, compact reasoning compared to direct GRPO training from preliminary SFT baseline.

2.8. Cross-Dataset Generalization

Table 6 demonstrates MedCausalX's superior zero-shot transfer across specialized medical domains. The model achieves an average accuracy of 57.4%. Gains are particularly notable on domain-specific tasks, including OVQA in ophthalmology with an improvement of 8.1% and Skin VQA in dermatology with an improvement of 9.9%. These results suggest that causal reasoning mechanisms, such as anatomical localization, pathology identification, and counterfactual verification, provide structural inductive biases that generalize effectively across modalities and anatomical regions. In contrast, reasoning-enhanced methods like Med-R1 and MedVLM-R1 achieve only modest improvements over standard models such as InternVL-2, indicating that chain-of-thought approaches lack sufficient compositional structure for reliable domain generalization.

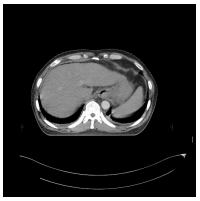
2.9. Training Dynamics Analysis

Figure 3 illustrates MedCausalX's training dynamics under on-policy optimization. The reward curves show steady improvement for both configurations, with MedCausalX (Full) benefiting from off-policy initialization to start higher. While the w/o Off-policy baseline gradually catches up, the Full version maintains a consistent advantage, reaching 0.845 versus 0.791. Response lengths reveal distinct trajectories: Full maintains compact reasoning around 850 tokens, reflecting DPO's pre-established efficiency. In contrast, the w/o Off-policy baseline starts at 919 tokens and stabilizes around 870 tokens as GRPO converges. These patterns confirm that off-policy training provides immediate efficiency gains, while GRPO alone requires longer optimization to achieve comparable compactness.

Prompt Templates

```
{
  "system_prompt": {
    "role": "system",
    "content": "You are an expert medical image analysis assistant with causal reasoning capabilities. Your task is to analyze medical images through structured causal decomposition following the A→P→Y (Anatomy→Pathology→Diagnosis) framework.

    1. Task Description: Examine medical images using causal factorization. Begin with precise anatomical localization (A), proceed to pathological characterization (P), and conclude with diagnostic inference (Y). Each stage must be causally grounded in the preceding observations.

    2. Reasoning Protocol: Employ two-stage reflective reasoning:
       - Stage 1 (<causal>): Generate preliminary causal analysis by systematically identifying anatomical structures, describing their spatial relationships, and linking visual features to pathological concepts. Provide bounding box coordinates [x1, y1, x2, y2] in normalized 0-1 range for all anatomical references.
       - Stage 2 (<verify>): Critically evaluate the initial causal chain for spurious correlations or logical inconsistencies. Refine anatomical localization, verify pathological characterizations against visual evidence, and correct any causal disruptions in the reasoning trajectory.

    3. Causal Chain Construction: For each reasoning stage, explicitly state:
       - Anatomical Observation: \"The <ref>structure_name</ref> <box>[x1,y1,x2,y2]</box> exhibits [specific features]\"
       - Pathological Inference: \"Based on [anatomical evidence], the pathological manifestation indicates [finding]\"
       - Diagnostic Conclusion: \"The causal chain establishes: [A→P→Y relationship]\"

    4. Output Format: Structure your response using reflective tokens:
       <causal>
       [Preliminary anatomical localization and initial pathological characterization with detailed spatial annotations]
       <verify>
       [Critical evaluation of causal dependencies, refinement of spatial localization, and correction of spurious associations]

    5. Causal Verification Principles:
       - Reject diagnoses based solely on global image statistics without anatomical grounding
       - Distinguish genuine causal dependencies from confounding correlations
       - Ensure spatial consistency between anatomical claims and bounding box coordinates
       - Validate that pathological characterizations logically follow from observed anatomical features"
  },


  "user_prompt": {
    "role": "user",
    "content": [
      {
        "type": "image",
        "image": " /path/to/medical_image.jpg "
      },
      {
        "type": "text",
        "text": "What pathological findings are present in this image?"
      }
    ]
  }
}
```

Figure 4. Structured prompt template for MedCausalX showing system instructions for pathological analysis approach and output format specifications enabling causal chain generation.



Case 1

Q: How many lungs are shown in this image?

<causal> This is a frontal chest radiograph displaying thoracic anatomy. I observe bilateral radiolucent lung fields extending from the apices to the diaphragmatic surfaces. The <ref>left lung</ref> <box>[0.52, 0.25, 0.85, 0.68]</box> demonstrates characteristic vascular markings extending throughout the pulmonary parenchyma with preserved aeration. The cardiac silhouette occupies the expected left-sided position. Moving to the contralateral side, the <ref>right lung</ref> <box>[0.15, 0.28, 0.48, 0.72]</box> exhibits similar radiolucency with clearly demarcated costophrenic angles and hemidiaphragm interfaces. Both lung fields maintain symmetric expansion patterns without evidence of consolidation, atelectasis, or pleural effusion. The mediastinal contours remain centered with preserved hilar structures bilaterally.

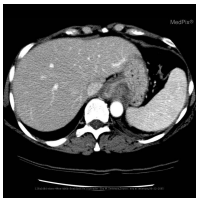
Step1

<verify> I need to confirm accurate enumeration and spatial boundary verification. Systematic examination validates two distinct anatomical units: the right pulmonary structure demonstrating three-lobed configuration and the left pulmonary structure with two-lobed morphology, both separated by the central mediastinum. The <ref>right hemithorax</ref> <box>[0.10, 0.22, 0.50, 0.75]</box> contains one complete lung unit with visible fissural margins, while the <ref>left hemithorax</ref> <box>[0.50, 0.20, 0.88, 0.70]</box> contains the second complete lung unit. No supernumerary pulmonary structures, accessory lobes, or congenital variants are present. Anatomical assessment identifies the bilateral thoracic cavities, and pathological evaluation demonstrates normally aerated pulmonary parenchyma, confirming the diagnostic conclusion of exactly two lungs.

Answer: 2

Step2

Figure 5. Chest radiograph demonstrating bilateral lung detection where MedCausalX accurately quantifies two distinct pulmonary structures through structured anatomical enumeration and spatial boundary verification.



Case 2

Q: Does this plane of section include the gastric cardia?

<causal> This axial contrast-enhanced CT section traverses the upper abdomen at the level of the gastroesophageal junction. I identify the
<ref>gastric fundus</ref> <box>[0.38, 0.35, 0.62, 0.68]</box>
as a gas-filled hollow viscus structure in the left upper quadrant with characteristic mucosal folds. Adjacent to this, at the anatomical transition zone where the esophagus penetrates the diaphragmatic hiatus, I observe the
<ref>gastric cardia region</ref> <box>[0.42, 0.38, 0.58, 0.52]</box>
demonstrating the proximal gastric pouch immediately inferior to the esophagogastric junction. The cardia represents the anatomical segment receiving esophageal contents, positioned at the gastric inlet where stratified squamous epithelium transitions to columnar mucosa. The imaging plane clearly captures this critical anatomical landmark with preserved wall thickness and no evidence of hiatal hernia displacement.

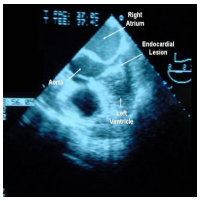
Step1

<verify> I need to confirm definitive cardia inclusion within this imaging section. Careful and systematic evaluation of the relevant anatomical structures demonstrates their expected spatial relationships and continuity. The distal esophagus descends smoothly through the esophageal hiatus, maintaining its normal caliber and orientation, while the gastroesophageal junction clearly marks the Z-line transition, delineating the boundary between esophageal and gastric mucosa. Immediately caudal lies the
<ref>proximal gastric cardia</ref> <box>[0.40, 0.42, 0.60, 0.56]</box>
before the fundic expansion. The hepatogastric ligament insertion and left gastric artery origin provide additional confirmation of cardia localization. The causal chain establishes that anatomical localization identifies the gastroesophageal transition zone, pathological characterization confirms intact mucosal architecture, and the diagnostic conclusion verifies the presence of the cardia within this section.

Answer: yes

Step2

Figure 6. Abdominal examination section validating gastric cardia identification through causal anatomical landmark detection, confirming presence via systematic spatial relationship analysis with esophagogastric junction.



Case 3

Q: What is the location of the echodense image seen in the echocardiogram?

<causal> This echocardiographic four-chamber view displays cardiac anatomy with clearly demarcated chamber boundaries and septal structures. I observe the <ref>right atrium</ref> <box>[0.28, 0.15, 0.58, 0.42]</box> in the superior right portion with the <ref>left ventricle</ref> <box>[0.42, 0.52, 0.75, 0.85]</box> occupying the inferior left quadrant. Within the cardiac septum, specifically along the <ref>lower intraventricular septum</ref> <box>[0.48, 0.58, 0.62, 0.78]</box>, I identify an echodense mass exhibiting irregular borders and heterogeneous echogenicity. This <ref>endocardial lesion</ref> <box>[0.52, 0.64, 0.58, 0.72]</box> projects into the left ventricular cavity with attachment to the septal endocardial surface, consistent with a fixed intracardiac mass rather than mobile thrombus. The lesion's position adjacent to rather than within the upper septum distinguishes it from membranous septal pathology.

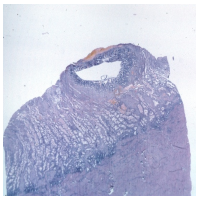
Step1

<verify> I need to precisely confirm the lesion's septal position relative to anatomical landmarks. Systematic evaluation establishes that the mass originates from the <ref>inferior interventricular septum</ref> <box>[0.46, 0.60, 0.64, 0.80]</box> below the mid-septal plane, specifically in the apical-to-mid ventricular region. The <ref>aortic root</ref> <box>[0.38, 0.32, 0.52, 0.48]</box> and mitral valve apparatus serve as superior reference points, confirming the lesion's distance from the upper septal structures. The mass demonstrates direct contact with the <ref>lower septal myocardium</ref> <box>[0.50, 0.62, 0.60, 0.76]</box> without extension to the superior membranous region. The causal chain validates that anatomical localization identifies the lower interventricular septum, pathological characterization reveals an endocardial-based echodense mass, and the diagnostic conclusion confirms the lesion's proximity to the lower, rather than the upper, septum.

Answer: Adjacent to the lower intraventricular septum

Step2

Figure 7. Echocardiographic image demonstrating successful causal localization of endocardial lesion positioned adjacent to lower intraventricular septum through structured anatomical-pathological reasoning chain.



Case 4

Q: Is uterus present?

<causal> This axial pelvic CT section demonstrates female reproductive anatomy within the true pelvis. I identify a midline soft tissue structure positioned between the <ref>urinary bladder</ref> <box>[0.35, 0.32, 0.65, 0.58]</box> anteriorly and the <ref>rectosigmoid colon</ref> <box>[0.38, 0.65, 0.62, 0.85]</box> posteriorly. This intermediate structure exhibits characteristic morphology of the <ref>uterine corpus</ref> <box>[0.42, 0.45, 0.58, 0.62]</box> with homogeneous myometrial density and preserved endometrial cavity echogenicity. The organ demonstrates typical pear-shaped configuration in the coronal plane with smooth serosal surfaces. Bilateral parametrial soft tissues and vascular pedicles provide additional anatomical confirmation. The uterus maintains normal anteverted position without evidence of surgical absence, congenital agenesis, or displacement by pelvic masses.

Step1

<verify> I need to definitively exclude confounding structures and confirm uterine identity. Systematic evaluation eliminates alternative diagnoses: the structure's location in the <ref>vesicouterine space</ref> <box>[0.40, 0.40, 0.60, 0.50]</box> excludes bowel loops; its homogeneous density distinguishes it from vascular structures; and its characteristic shape eliminates ovarian or adnexal masses. The <ref>cervicouterine junction</ref> <box>[0.44, 0.58, 0.56, 0.64]</box> transitions into the superior vaginal fornix inferiorly, confirming anatomical continuity. Bilateral <ref>broad ligament attachments</ref> <box>[0.25, 0.48, 0.75, 0.58]</box> extend laterally toward pelvic sidewalls. The causal chain establishes that anatomical localization identifies a midline pelvic structure, pathological characterization reveals typical uterine morphology and density, and the diagnostic conclusion confirms the presence of the uterus unambiguously.

Answer: yes

Step2

Figure 8. Pelvic CT scan validating binary anatomical detection where MedCausalX correctly confirms uterine presence through causal verification eliminating confounding structures.

References

[1] Shuai Bai and et al. Qwen2.5-vl: Technical report. https://arxiv.org/abs/2502.13923, 2025. arXiv preprint arXiv:2502.13923. 1

[2] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 3, 4

[3] Xuehai He, Zhuo Cai, Wenlan Wei, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathological visual question answering. *arXiv preprint arXiv:2010.12435*, 2020. 1

[4] 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. 4

[5] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019. 1

[6] Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, Yuheng Li, Konstantinos Psounis, and Xiaofeng Yang. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *IEEE Transactions on Medical Imaging*, 2026. 3, 4

[7] 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. 1

[8] 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. 4

[9] 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. 1

[10] Haotian Liu et al. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2024. 3

[11] 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. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 337–347. Springer, 2025. 3, 4

[12] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023. 1

[13] 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. 1

[14] Lehan Wang, Haonan Wang, Honglong Yang, Jiaji Mao, Zehong Yang, Jun Shen, and Xiaomeng Li. Interpretable bilingual multimodal large language model for diverse biomedical tasks. *arXiv preprint arXiv:2410.18387*, 2024. 3

[15] Jin Ye, Junlong Cheng, Jianpin Chen, Zhongying Deng, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiang, et al. Sa-med2d-20m dataset: Segment anything in 2d medical imaging with 20 million masks. *arXiv preprint arXiv:2311.11969*, 2023. 1