THINK BEFORE YOU LOCATE: CLINICAL CHAIN-OF-THOUGHT SUPERVISION FOR SEQUENTIAL MEDICAL IMAGE GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Clinicians rarely localize a finding from one image alone. They compare current and prior scans, follow structures across slices, and match regions across views or modalities. Yet vision-language models for sequential medical grounding usually learn only from bounding boxes. Boxes show where the target is, but they do not explain why that region in that image answers the question. We study whether explicit, image-grounded reasoning can improve localization, and which tasks benefit from it. We introduce MEDCHAIN-R1, whose core is MedChain, a corpus of 182,573 filtered rationales spanning eight sequential grounding tasks and ten imaging modalities. A medical VLM (FlemingVL-38B) first drafts a caption-like rationale, and a multimodal refiner (Qwen3.5-397B-A17B) then re-examines the images, question and annotated box to write a box-conditioned explanation with explicit cross-image reasoning. We fine-tune Qwen3.5 on these rationales with deliberative chain-of-thought supervision (dCoT-SFT). Across 4B, 9B, and 27B models, dCoT-SFT sets a new state of the art on MedSG-Bench. The 27B model reaches 77.24 mIoU and 85.20 Acc@0.5, improving over the strongest prior model by 4.69 and 5.49 points. In a paired comparison using fixed 4B checkpoints, dCoT-SFT improves over answer-only SFT by 3.56 mIoU (95% CI 3.13–4.00). The gains mainly come from change detection and object tracking, while performance drops on three appearance-matching tasks. Replacing a model's own rationale with a rationale from another task reduces mIoU by 12.2 points, showing that localization depends strongly on the reasoning context. Finally, adding GRPO with an IoU reward after dCoT-SFT gives only small and task-dependent gains, which we analyze through controlled multi-seed development experiments.

1 INTRODUCTION

A radiologist reading a follow-up chest CT (Computed Tomography) rarely cares about a nodule in isolation; the key question is whether it has changed since the prior study. The same reasoning is needed to track a lesion across slices, match regions between FLAIR and diffusion-weighted MRI, or find the same mass across mammographic views. These tasks share one core ability: localizing a target by comparing a sequence of images. Medical visual grounding has advanced quickly (Zou et al., 2024; Boecking et al., 2022; Bannur et al., 2023), but most models still focus on single images. MedSG-Bench (Yue et al., 2026) made this gap measurable with 9,630 questions across eight tasks and ten modalities. Most general-purpose and medical multimodal LLMs score below 20 mIoU, while MedSeq-Grounder reaches 72.55 mIoU after instruction tuning on 188K sequential examples.

Box-only supervision leaves the key reasoning implicit. To answer *what changed between these two scans?*, a model must match anatomy across images, compare the regions, ignore differences from windowing or positioning, and then place the box. An answer-only objective rewards the final coordinates without teaching which comparison supports them. Prior work in natural-image grounding and medical visual question answering shows that intermediate reasoning can improve accuracy beyond supervised baselines (Huang et al., 2026; Bai et al., 2025c; Pan et al., 2025; Lai et al., 2026; Xu et al., 2025). Sequential medical grounding has not yet used this approach because no reasoning corpus exists for the setting, and text-only reasoning models cannot inspect the images they need to explain.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

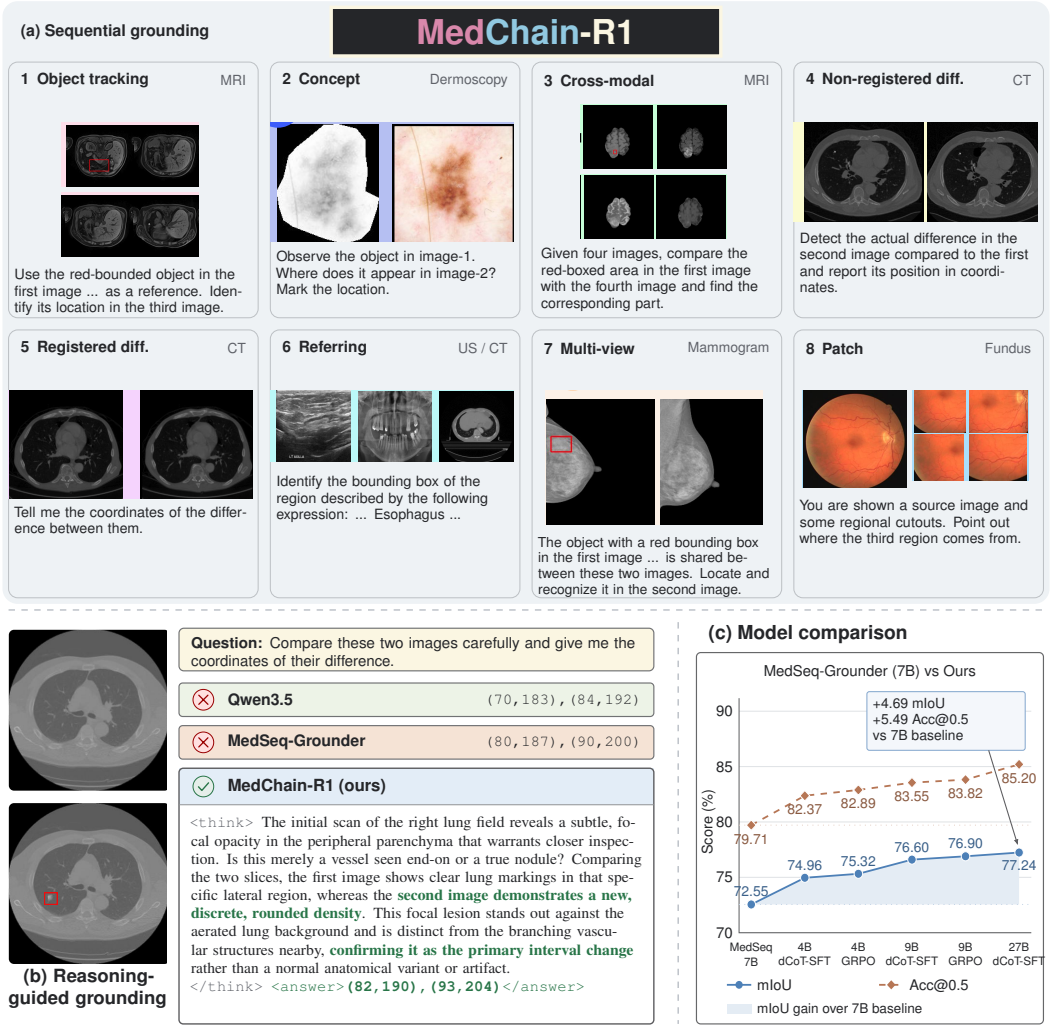


Figure 1: **Sequential medical grounding with reasoning supervision.** (a) The eight MedSG-Bench tasks, grouped into image difference grounding (IDG) and image consistency grounding (ICG). (b) A registered-difference example: zero-shot Qwen3.5 and MedSeq-Grounder return imprecise boxes, whereas MEDCHAIN-R1 first identifies the new nodule and then localizes it. (c) MedSG-Bench mIoU and Acc@0.5 of MedSeq-Grounder (7B) and MEDCHAIN-R1 at 4B and 9B (dCoT-SFT and dCoT-SFT+GRPO) and 27B (dCoT-SFT).

We address both problems with MEDCHAIN-R1. We build its corpus, MedChain, with a two-stage multimodal pipeline (Figure 2). First, a medical VLM drafts a preliminary rationale from the image sequence; on inspection, these drafts read like captions and rarely state the cross-image comparison the task requires. Then, a large multimodal model re-examines the original images, question, task context, and annotated box, and writes a short explanation of why the boxed region answers the question. The refiner keeps direct access to the images, which matters when the decisive evidence is a cross-image relation that a caption omits. We use the box as privileged information only during offline corpus construction. Each rationale therefore explains a known answer rather than recording an independent search process. We make this choice deliberately and directly test how the trained model uses its reasoning at inference time. We fine-tune Qwen3.5 (Team, 2026) at 4B, 9B, and 27B parameters on this think-then-answer format, which we call dCoT-SFT. We then test whether GRPO (Shao et al., 2024) with localization rewards provides additional gains.

Our experiments give four main findings:

1. **Reasoning supervision sets a new SOTA.** dCoT-SFT outperforms MedSeq-Grounder at every model scale, reaching 74.96, 76.60, and 77.24 mIoU at 4B, 9B, and 27B, so average performance rises monotonically with model size (Table 1).
2. **The gains depend strongly on the task.** Compared with an answer-only 4B model, dCoT-SFT improves overall mIoU by 3.56 points (95% CI 3.13–4.00). It gains 14.7–16.1 points on difference grounding and 6.5 points on object tracking, but loses 2.2–3.2 points on multi-view, patch, and referring grounding (Table 2).
3. **Predictions depend on the reasoning context.** Forcing the dCoT-SFT model to answer immediately reduces mIoU by 2.9 points. Conditioning it on a rationale from another task reduces mIoU by 12.2 points.
4. **Reward optimization adds little after dCoT-SFT.** GRPO with an IoU+format reward improves benchmark mIoU by only 0.37 points and hurts concept grounding. In controlled three-seed development runs, it changes aggregate mIoU by less than 0.1 points, with a consistent gain only on multi-view grounding (Table 3).

Our contributions are: (i) a two-stage multimodal rationale pipeline in which a medical VLM describes the images and a multimodal refiner re-inspects the full sequence to write the cross-image reasoning; (ii) MedChain, to our knowledge the first reasoning corpus for sequential medical grounding; (iii) dCoT-SFT models at three scales that set a new state of the art; and (iv) a controlled analysis of when reasoning helps, whether predictions depend on it, and what GRPO adds on top. We document MedChain with row- and ID-level accounting, a train–test image overlap audit, and a single-expert rationale audit. We report uncertainty throughout using paired example-level tests.

2 Related Work

Medical visual grounding. Medical visual grounding supports interpretable vision-language models by linking text to specific image regions (Zou et al., 2024; Boecking et al., 2022; Bannur et al., 2023). Most prior work localizes findings in a single chest radiograph or pathology image (Chen et al., 2024d; Bannur et al., 2024), while large instruction-tuned medical models extend this capability across modalities (Li et al., 2023; Chen et al., 2024a; Sellergren et al., 2025). MedSG-Bench (Yue et al., 2026) introduced sequential grounding over longitudinal, multi-view, and cross-modal image sets, together with MedSeq-Grounder, which uses answer-only supervision. We use the same benchmark and study what explicit reasoning supervision adds.

Reasoning supervision and reinforcement learning for grounding. Chain-of-thought prompting (Wei et al., 2022) and multimodal instruction tuning (Liu et al., 2023; Dai et al., 2023) showed that explicit intermediate reasoning can improve complex visual tasks. DeepSeek-R1 (Guo et al., 2025) popularized a cold-start reasoning SFT stage followed by GRPO (Shao et al., 2024) with verifiable rewards, while DAPO (Yu et al., 2026) improved training stability. Vision-R1 (Huang et al., 2026) and UniVG-R1 (Bai et al., 2025c) applied this approach to natural-image grounding. MedVLM-R1 (Pan et al., 2025), Med-R1 (Lai et al., 2026), and MedGround-R1 (Xu et al., 2025) extended it to medical VQA and single-image medical grounding. We study the multi-image setting and find that reasoning SFT provides most of the gain, while IoU-rewarded GRPO adds little.

Rationale distillation from multimodal teachers. Prior work commonly distills rationales from large models into smaller models to scale reasoning supervision (Guo et al., 2025; Wei et al., 2022). In medicine, the main challenge is keeping fluent explanations tied to visual evidence. Our refiner is multimodal: it inspects the original images rather than relying only on the teacher's text. We also condition the refiner on the annotated box. This answer conditioning improves alignment with the target, but it can also produce post-hoc rationalization. We therefore test whether the trained model actually relies on its reasoning in Section 5.2. Appendix H discusses additional related work.

3 MedChain-R1: Reasoning Supervision for Sequential Grounding

We present MedChain-R1, which introduces explicit image-grounded reasoning supervision for sequential medical grounding through the MedChain rationale corpus, as illustrated in Figure 2. We use these rationales to train models with dCoT-SFT, then test whether GRPO adds further benefit beyond supervised reasoning.

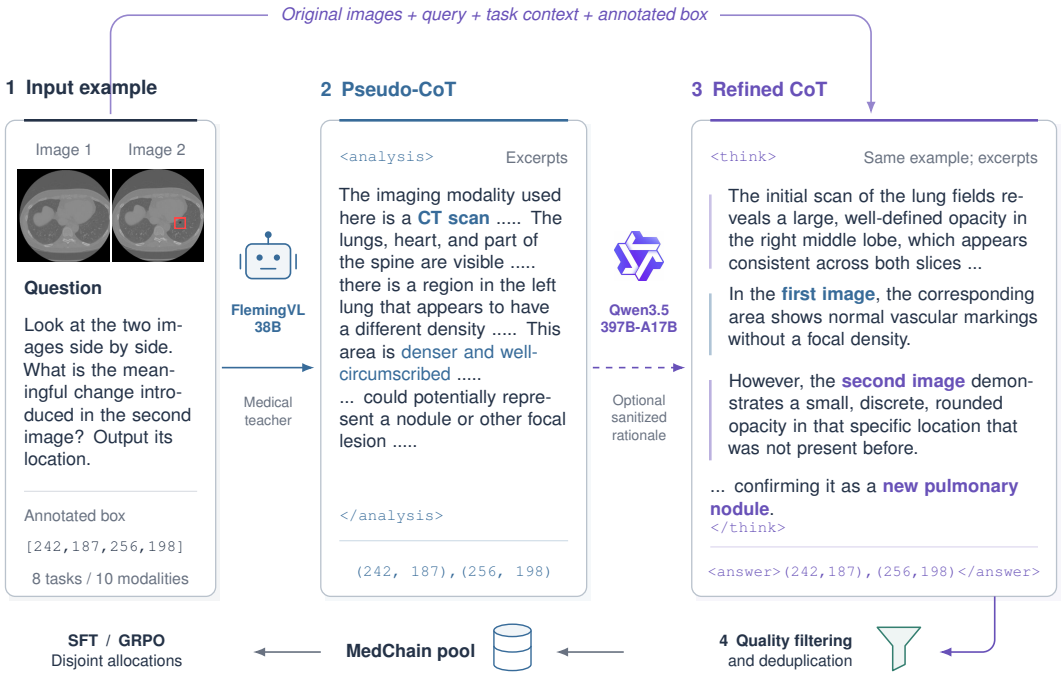


Figure 2: **MedChain construction.** (1) Each training example contains an image sequence, a clinical question and an annotated box. (2) FlemingVL-38B drafts a pseudo-CoT from the images. (3) Multimodal Qwen3.5-397B-A17B receives the original images, question, task context, annotated box and an optional sanitized pseudo-CoT, and writes an image-grounded explanation in `<think>`/`<answer>` format. (4) Outputs are filtered and deduplicated into the MedChain pool, which is split into disjoint SFT and GRPO allocations.

Task Description. Given a sequence of N medical images $\mathcal{I} = \{I_1, \ldots, I_N\}$ and a clinical query q, we predict a box $\hat{b} = (x_1, y_1, x_2, y_2)$ that localizes the target in the correct image $I_k \in \mathcal{I}$. MedSG-Bench divides its eight tasks into two paradigms (Yue et al., 2026). In *image difference grounding* (IDG: registered and non-registered difference), the prompt gives the target image and the model must localize what changed. In *image consistency grounding* (ICG: multi-view, object tracking, concept, patch, cross-modal, and referring grounding), the model must find the same structure across changes in view, slice, modality, or scale, and may also need to identify the target image. Both paradigms require cross-image comparison, while ICG additionally requires image selection.

3.1 MedChain: Image-Grounded Rationales

Stage 1: Medical pseudo-CoT. Text-only reasoning models cannot inspect CT, MRI, ultrasound or pathology images. We therefore start from a medical VLM. After comparing six candidate teachers on 100 sampled instances each (Appendix A.1), we use FlemingVL-38B (Shu et al., 2025) to produce a preliminary rationale $\tilde{c} = \mathcal{F}(\mathcal{I}, q, b^*)$ describing the modality, anatomy and visible content of each image, where b^* is the annotated box. FlemingVL-38B acts mainly as a medical captioner: although prompted to relate the images, it describes what each image contains and rarely states how the images relate.

Stage 2: Multi-modal refinement. The Stage 1 rationales capture visual content but read as captions: they describe each image without the cross-image comparison, self-checks or reasoning steps that the target requires. Stage 2 supplies this reasoning. A multimodal refiner, Qwen3.5-397B-A17B (Team, 2026), receives the original images together with the question, task context t, annotated box and an optional sanitized $\tilde{c}$, and writes the final rationale $c^* = \mathcal{Q}(\mathcal{I}, q, t, b^*, \tilde{c})$. The refiner is instructed to

Algorithm 1 MedChain construction via medical modality bridging.

Require: examples $\mathcal{D} = \{(\mathcal{I}_i, q_i, t_i, b_i^*)\}_{i=1}^{M}$; teacher $\mathcal{F}$ (FlemingVL-38B); refiner $\mathcal{Q}$ (Qwen3.5-397B-A17B); quality threshold τ

$\mathcal{C} \leftarrow \emptyset$; $\mathcal{E} \leftarrow \emptyset$ ▷ accepted rationales; failures
for each $(\mathcal{I}_i, q_i, t_i, b_i^*) \in \mathcal{D}$ **do**
  $\tilde{c}_i \leftarrow \mathcal{F}(\mathcal{I}_i, q_i, b_i^*)$ ▷ Stage 1: medical pseudo-CoT
  if $\tilde{c}_i$ is malformed **or** the context exceeds 8K tokens **then** $\mathcal{E} \leftarrow \mathcal{E} \cup \{i\}$; **continue**
  $c_i^* \leftarrow \mathcal{Q}(\mathcal{I}_i, q_i, t_i, b_i^*, \tilde{c}_i)$ ▷ Stage 2: multimodal refinement
  if c_i^* is malformed **or** its answer box $\neq b_i^*$ **then** $\mathcal{E} \leftarrow \mathcal{E} \cup \{i\}$; **continue**
  if quality$(c_i^*) \geq \tau$ **then** $\mathcal{C} \leftarrow \mathcal{C} \cup \{(\mathcal{I}_i, q_i, c_i^*, b_i^*)\}$
end for
Reprocess $\mathcal{E}$ with adjusted context splitting and merge into $\mathcal{C}$
Deduplicate $\mathcal{C}$ (182,573 IDs); split into disjoint $\mathcal{C}_{\text{GRPO}}$ (32,000) and $\mathcal{C}_{\text{SFT}}$ (150,573)
return $\mathcal{C}_{\text{SFT}}$, $\mathcal{C}_{\text{GRPO}}$

Table 1: **MedSG-Bench results.** Each cell shows mIoU (top) and Acc@0.5 (bottom), in percent. IDG: registered (RDG) and non-registered (NRDG) difference grounding. ICG: multi-view (MV), object tracking (OT), concept (VCG), patch (VPG), cross-modal (CMG) and referring (RG) grounding. **Bold**: our results that exceed MedSeq-Grounder, the strongest prior model, on the same metric; small numbers in parentheses give the difference from MedSeq-Grounder (green: higher, red: lower). Results for all baselines are in Table 10.

	Model	Size	RDG	NRDG	MV	OT	VCG	VPG	CMG	RG	Avg.
General	Qwen2.5-VL	72B	4.37 2.60	3.46 0.80	7.22 2.78	13.11 7.70	10.33 3.50	26.45 6.30	16.32 7.00	20.19 14.10	13.35 6.12
	InternVL3	78B	0.20 0.00	0.53 0.00	6.35 2.43	13.03 8.00	3.57 0.90	11.81 2.50	3.34 0.85	12.76 8.10	6.44 2.90
	LLaVA-OneVision	72B	2.58 0.80	2.87 0.90	11.74 1.39	9.61 2.30	10.95 3.30	32.38 20.30	16.24 5.40	15.43 6.68	13.21 5.18
	Migician	7B	15.26 7.80	14.49 6.10	18.16 7.84	21.38 14.90	14.23 7.20	28.87 13.70	21.41 12.15	25.30 18.02	20.29 11.39
	Qwen3.5 (zero-shot)	27B	0.00 0.00	0.00 0.00	8.48 4.88	28.39 24.50	0.25 0.00	2.50 0.20	0.25 0.00	0.82 0.00	4.42 3.15
Medical	HuatuoGPT-Vision	34B	1.44 0.00	2.15 0.00	9.41 1.65	13.25 8.30	6.43 0.70	14.53 1.40	10.60 2.60	8.60 1.75	8.57 2.09
	MedGemma	27B	0.01 0.00	0.06 0.00	5.29 2.26	22.26 16.10	2.27 0.00	3.11 0.30	2.92 0.35	5.35 0.61	4.94 2.14
	MedGRPO-7B-RL	7B	0.09 0.00	0.06 0.00	3.68 1.48	17.63 15.00	4.57 0.40	18.44 1.20	0.78 0.15	9.31 3.37	6.27 2.45
	MedSeq-Grounder	7B	83.29 93.20	83.72 94.10	55.03 60.19	62.10 67.20	74.11 82.60	85.25 98.80	78.77 82.75	60.43 65.59	72.55 79.71
Ours	MEDCHAIN-R1 (dCoT-SFT)	4B	82.86 (−0.43)	81.30 (−2.42)	54.01 (−1.02)	**64.77** (+2.67)	70.64 (−3.47)	**90.21** (+4.96)	**83.39** (+4.62)	**69.67** (+9.24)	**74.96** (+2.41)
			93.30 (+0.10)	91.20 (−2.90)	58.97 (−1.22)	**71.00** (+3.80)	78.80 (−3.80)	**99.70** (+0.90)	**87.00** (+4.25)	**79.28** (+13.69)	**82.37** (+2.66)
	MEDCHAIN-R1 (+GRPO)	4B	**83.82** (+0.53)	83.21 (−0.51)	54.80 (−0.23)	**67.90** (+5.80)	66.09 (−8.02)	**89.43** (+4.18)	**83.59** (+4.82)	**70.72** (+10.29)	**75.32** (+2.77)
			94.50 (+1.30)	92.70 (−1.40)	59.93 (−0.26)	**76.20** (+9.00)	72.10 (−10.50)	**99.80** (+1.00)	**87.55** (+4.80)	**80.30** (+14.71)	**82.89** (+3.18)
	MEDCHAIN-R1 (dCoT-SFT)	9B	**84.00** (+0.71)	83.02 (−0.70)	**55.87** (+0.84)	**67.46** (+5.36)	**74.64** (+0.53)	**91.22** (+5.97)	**84.16** (+5.39)	**70.77** (+10.34)	**76.60** (+4.05)
			93.10 (−0.10)	91.60 (−2.50)	**60.98** (+0.79)	**75.00** (+7.80)	**83.20** (+0.60)	**99.90** (+1.10)	**87.40** (+4.65)	**78.95** (+13.36)	**83.55** (+3.84)
	MEDCHAIN-R1 (+GRPO)	9B	**84.16** (+0.87)	83.12 (−0.60)	**56.98** (+1.95)	**67.65** (+5.55)	**74.80** (+0.69)	**91.22** (+5.97)	**84.82** (+6.05)	**70.53** (+10.10)	**76.90** (+4.35)
			93.20 (0.00)	91.40 (−2.70)	**62.02** (+1.83)	**74.90** (+7.70)	**82.80** (+0.20)	**99.80** (+1.00)	**88.25** (+5.50)	**79.22** (+13.63)	**83.82** (+4.11)
	MEDCHAIN-R1 (dCoT-SFT)	27B	81.87 (−1.42)	80.97 (−2.75)	**60.14** (+5.11)	**69.53** (+7.43)	**74.64** (+0.53)	**91.87** (+6.62)	**84.24** (+5.47)	**72.51** (+12.08)	**77.24** (+4.69)
			93.00 (−0.20)	91.60 (−2.50)	**66.72** (+6.53)	**78.30** (+11.10)	**83.30** (+0.70)	**99.90** (+1.10)	**88.15** (+5.40)	**81.98** (+16.39)	**85.20** (+5.49)

justify the boxed region from visible evidence, contrast it with nearby alternatives, include one or two brief self-checks, and avoid mentioning that the box was given (prompts in Appendix B). Because it inspects the images directly, c^* does not depend only on Stage 1 text. Because it sees b^*, c^* is an explanation of the annotated answer, not a faithful record of how any model found it.

Filtering and accounting. We discard outputs that are malformed, exceed the Stage 1 context limit, or produce a final box that does not match b^*. We reprocess failed cases using adjusted context splitting (Algorithm 1). After quality filtering and deduplication, MedChain contains 182,573 unique example IDs. We reserve 32,000 IDs for reinforcement learning, with 4,000 per task using seed 42. The remaining 150,573 IDs form the SFT split, released as 152,396 rows. We also audit train–test overlap using encoded and decoded-pixel hashes. We find no complete test image sequence and no source-series-plus-box pair in the SFT split. However, 147 test examples share at least one individual image with the training data (Appendix A.5), so we also report results after excluding these examples.

Rationale audit. One medical expert reviews 2,000 rationales for clinical plausibility, consistency with the image sequence and task, and agreement with the target. The expert accepts 1,965 rationales (98.25%; Wilson 95% CI 97.58–98.74). The remaining 35 contain minor-to-moderate reasoning errors, although all have correct boxes. This audit uses a single rater and should not be treated as multi-specialist clinical validation (Table 5).

3.2 Training Protocol

dCoT-SFT. We fine-tune Qwen3.5 models at 4B, 9B, and 27B with deliberative chain-of-thought supervised fine-tuning (dCoT-SFT), i.e. full-parameter supervised learning on MedChain rationales. We use the format `<think>` c^* `</think><answer>` b^* `</answer>` (Guo et al., 2025), so the model learns to state its cross-image reasoning before predicting the coordinates. We train with LLaMA-Factory (Zheng et al., 2024) using AdamW, a cosine learning-rate schedule, a learning rate of 5×10^{-6}, and an effective batch size of 64. We select checkpoints using minimum validation loss (Appendix A.8).

GRPO on top of dCoT-SFT. We then test whether reward optimization adds further gains by training the 4B and 9B dCoT-SFT models with GRPO (Shao et al., 2024). For each prompt, the policy samples $G = 8$ completions. We assign each completion the reward $r = r_{\text{IoU}} + r_{\text{fmt}}$, where r_{IoU} measures the IoU between the predicted box and b^*, and r_{fmt} checks the output format. We compute each advantage relative to the other completions sampled for the same prompt. We train on the 32,000 held-out GRPO examples with KL coefficient $\beta = 0.04$, clipping $\epsilon = 0.2$, and learning rate 10^{-6} for up to 3,000 steps, selecting each checkpoint on a held-out development set (Appendix C.3).

4 Experiments

Setup. We evaluate on all 9,630 MedSG-Bench test questions (Yue et al., 2026) and report mIoU and Acc@0.5, where Acc@0.5 is the percentage of predictions with IoU ≥ 0.5. We evaluate all baselines zero-shot using one unified prompt (Appendix A.6). These include general-purpose and medical MLLMs from 1.5B to 78B parameters (full list in Appendix C), the RL-trained MedGRPO-7B-RL, and the supervised SOTA model, MedSeq-Grounder (Yue et al., 2026).

Statistical analysis. Table 1 reports representative results, while Appendix C contains the full comparison. For paired analyses, we match predictions by example ID and use 100,000 example-level bootstrap resamples (10,000 for the reasoning-context controls) together with sign-flip tests. The resulting intervals measure uncertainty from test-set sampling, not variation across training runs.

5 Experimental Findings

5.1 Overall Findings

Zero-shot models struggle with sequential grounding. No zero-shot model exceeds 21 mIoU (Table 1). Larger general models do not perform better: InternVL3-78B scores below its 8B version, while the Qwen3.5 base models, including 27B, score below 5 mIoU under pixel-coordinate scoring, largely because they output degenerate or copied boxes (Appendix G). The coordinate convention explains only part of this gap: in a separate thinking-off rerun, reading the Qwen3.5 boxes as 0–1000 normalized coordinates raises overall mIoU to at most 13.30 (27B), still far below fine-tuned models (Appendix A.7). Medical pretraining also provides little benefit. MedGRPO-7B-RL, which uses RL on adjacent surgical video frames, transfers poorly and reaches only 6.27 mIoU. Scale,

medical knowledge and nearby video supervision alone therefore do not teach the cross-image correspondences that sequential grounding requires.

Reasoning supervision surpasses the answer-only SOTA. Our dCoT-SFT models outperform MedSeq-Grounder at every scale. They improve mIoU by 2.41 points at 4B, 4.05 at 9B, and 4.69 at 27B, with corresponding Acc@0.5 gains of 2.66, 3.84, and 5.49 points. At 27B, we see the largest gains on referring (+12.1 mIoU), object tracking (+7.4), patch (+6.6), and cross-modal grounding (+5.5). Part of this margin reflects the backbone: our answer-only 4B model already exceeds MedSeq-Grounder on RG, VPG and CMG (Table 2), whereas reasoning supervision is what lifts the average above it (71.39 to 74.96 mIoU, versus 72.55). On difference grounding, our models perform similarly to MedSeq-Grounder, which remains strongest on NRDG. Since the two model families use different backbones and training data, we treat this comparison as descriptive. As a partial backbone control, we continue dCoT-SFT from the released MedSeq-Grounder checkpoint (Qwen2.5-VL-7B) and improve its mIoU/Acc@0.5 from 72.55/79.71 to 75.71/82.34 (Appendix C.2). A matched answer-only continuation is not available.

Scaling. Average performance rises monotonically with model size (74.96, 76.60 and 77.24 mIoU), but the effect differs across tasks. From 9B to 27B, ICG performance improves or remains stable, with gains of +4.3 mIoU on MV, +2.1 on OT, and +1.7 on RG. In contrast, both IDG tasks drop by about 2 points. However, the 27B model uses a different training schedule: six epochs, a smaller warmup, and no weight decay, and it reaches a higher validation loss (Appendix A.8). We therefore do not attribute this difference to model size alone.

5.2 When and How Does Reasoning Help?

Controlled comparison. The comparison with MedSeq-Grounder mixes differences in supervision, backbone, and training data. To isolate supervision as far as our runs allow, we compare fixed 4B checkpoints trained with answer-only SFT and dCoT-SFT on the same backbone for the same number of optimization steps (Table 2, top). All 9,630 predictions match exactly by example ID, task, and ground-truth box, with no parser failures.

Overall effect. dCoT-SFT improves mIoU by 3.56 points (95% CI 3.13–4.00) and Acc@0.5 by 2.06 points (1.36–2.75), with $p < 10^{-5}$ under both bootstrap and sign-flip tests. The result remains almost unchanged after excluding the 164 test examples that share an image or source series with the SFT data: $+3.61$ mIoU (3.17–4.04) and $+2.01$ Acc@0.5 (1.31–2.70) on the remaining 9,466 examples. The stable gain after this exclusion suggests that the improvement is not driven by the identified train–test overlap. The answer-only run was trained on fewer nodes, so its effective batch size, optimizer implementation and checkpoint-selection rule still differ, and neither run was repeated over seeds; we therefore treat this as a fixed-model comparison rather than a causal or multi-seed estimate. The answer-only model also scores below MedSeq-Grounder on difference grounding, which uses a different backbone, 188K training examples and its own schedule, so the absolute size of the IDG gain should be read with this in mind.

The effect depends strongly on the task. dCoT-SFT gains most when localization is defined by comparison: registered and non-registered difference grounding improve by 16.1 and 14.7 mIoU, while object tracking improves by 6.5. In contrast, performance drops on multi-view (-2.2), patch (-2.5), and referring grounding (-3.2). This split suggests that explicit reasoning is most useful when the model must identify a relation across images before it can localize the target. Patch and referring grounding may rely more on local appearance or text-to-region matching, where verbal reasoning adds less. Multi-view grounding also requires comparison, however, so its decline does not fit this explanation and remains unresolved. We therefore report the pattern as empirical rather than as a validated mechanism.

Predictions depend on the reasoning context. Because MedChain rationales are written with the answer in view, the model could learn to produce plausible reasoning while localizing independently of it. We test this with two inference-time interventions on the fixed 4B dCoT-SFT checkpoint (Table 2, bottom; Appendix C.4). Forcing an immediate answer (*no-think*) lowers mIoU from 74.96 to 72.02 (-2.93, 95% CI -3.29 to -2.58). Prefilling a deterministic rationale from another task lowers it further to 62.77 (-12.18, -12.75 to -11.60); all tests give $p < 10^{-4}$ with 10,000 paired resamples. The much larger drop under mismatched reasoning suggests that localization is sensitive not only to the presence of reasoning, but also to its task-specific content. The effect is

Table 2: **Controlled 4B analysis.** All rows use the same 9,630 test questions. Top: answer-only SFT versus dCoT-SFT checkpoints of Qwen3.5-4B, and their difference (Δ, computed from the rounded entries). Bottom: inference-time interventions on the same dCoT-SFT checkpoint. Cells show mIoU (top) and Acc@0.5 (bottom). Gains are shaded green and losses red.

Qwen3.5-4B	RDG	NRDG	MV	OT	VCG	VPG	CMG	RG	Overall
Answer-only SFT	66.76 81.00	66.64 81.40	56.21 63.24	58.27 66.40	69.03 78.40	92.66 99.70	80.83 85.10	72.85 83.47	71.39 80.31
dCoT-SFT	82.86 93.30	81.30 91.20	54.01 58.97	64.77 71.00	70.64 78.80	90.21 99.70	83.39 87.00	69.67 79.28	74.96 82.37
Δ mIoU	+16.10	+14.66	−2.20	+6.50	+1.61	−2.45	+2.56	−3.18	+3.56
dCoT-SFT, forced no-think	82.15 92.20	80.94 90.80	49.72 53.57	55.26 58.90	69.40 76.40	87.55 99.00	81.44 84.85	66.36 73.55	72.02 78.66
dCoT-SFT, cross-task rationale	65.19 73.30	80.16 89.60	39.55 41.55	46.74 50.60	65.12 72.40	86.03 96.80	65.78 70.00	56.88 63.02	62.77 68.93

largest on object tracking (-18.0), registered difference (-17.7), cross-modal (-17.6), and multi-view grounding (-14.5), and smallest on non-registered difference grounding (-1.1). Removing reasoning hurts most on object tracking (-9.5) and multi-view grounding (-4.3). Together, these results suggest that reasoning context matters most for tasks that must track or match structures across images. These interventions do not prove that the rationale faithfully records the model's internal computation, and the cross-task rationale itself introduces a distribution shift.

5.3 Does Reward Optimization Help Beyond dCoT-SFT?

We add GRPO as a second training stage and test whether reward optimization improves on reasoning supervision.

Benchmark effect. GRPO improves mIoU by 0.37 points (95% CI 0.05–0.69, $p = 0.025$), but the Acc@0.5 gain is not significant ($+0.52$, -0.04 to 1.08, $p = 0.075$) (Table 3). The average also hides opposite task effects: object tracking gains 3.1 mIoU and 5.2 Acc@0.5, while concept grounding loses 4.6 and 6.7, respectively (Figure 3). This pattern suggests that an overlap-based reward can improve localization on some tasks while weakening semantic identification on others. IoU rewards overlap regardless of whether the model identifies the correct finding, which may matter most for concept grounding, where the model must recognize a lesion or structure across very different images. We do not isolate this mechanism here.

Controlled development studies. We next run controlled GRPO experiments on a separate 1,385-example development set (Appendix C.7). An initial 3–5 point drop was an evaluation artifact: the GRPO export saved an image processor with a lower resolution limit. After restoring the original processor, a zero-learning-rate save-and-reload control exactly reproduces all SFT predictions. With matched preprocessing, direct GRPO at a fixed step changes aggregate mIoU by only $+0.06$ (SD 0.08) and Acc@0.5 by $+0.24$ (SD 0.15) across three seeds. The only consistent task-level gain is on multi-view grounding, with $+4.4$ mIoU and $+7.3$ Acc@0.5, although this task has only 59 examples. A task-balanced variant with a threshold-aware reward gives $+0.36$ mIoU and $+0.94$ Acc@0.5, but neither interval excludes zero, and the gain disappears after a prompt-formatting change.

Headroom without gains. We finally test whether the policy can already sample better answers than it normally returns. On 128 training prompts from the four hardest tasks, individual samples reach IoU ≥ 0.5 in 67.7% of cases on average, while at least one of eight samples succeeds for 84.4% of prompts. This 16.7-point gap shows that rollout groups often contain a better localization than the policy's typical output. However, all eight samples fail on 20 of the 128 prompts, and GRPO does not convert the available headroom into benchmark gains in our configurations. Possible reasons include weak within-group contrast on the hardest examples, large differences in starting difficulty across tasks (54–90 mIoU), and a reward that measures spatial overlap without semantic correctness. These remain hypotheses to test with alternative rewards.

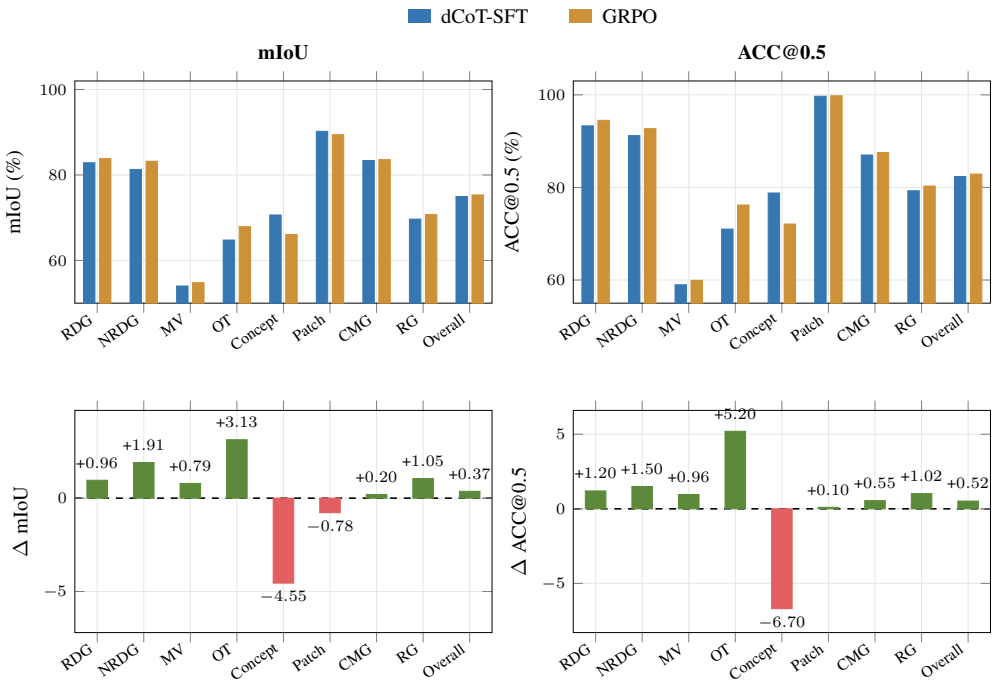


Figure 3: **Task-wise effect of GRPO after dCoT-SFT (4B, full test set).** Top: mIoU and Acc@0.5 of the fixed dCoT-SFT checkpoint and the GRPO checkpoint. Bottom: per-task differences. GRPO improves most localization-style tasks and object tracking in particular, but lowers concept grounding by 4.55 mIoU and 6.70 Acc@0.5. Only concept grounding declines substantially under GRPO.

Table 3: **GRPO after dCoT-SFT (4B; rows marked 9B use the 9B model).** Paired differences from the dCoT-SFT starting point in percentage points, with 95% bootstrap CIs over examples or the standard deviation over training seeds. Development results use a separate 1,385-example set and are exploratory. The 9B all-task row is a point estimate from example-weighted means, without a paired interval.

Evaluation	**GRPO variant**	**Δ mIoU**	**Δ Acc@0.5**
Test, all tasks (n=9,630)	IoU+format	+0.37 [0.05, 0.69]	+0.52 [−0.04, 1.08]
Test, concept task	IoU+format	−4.55	−6.70
Test 9B, all tasks	IoU+format	+0.30	+0.27
Test 9B, multi-view	IoU+format	+1.11 [0.39, 1.83]	+1.05 [−0.44, 2.53]
Test 9B, cross-modal	IoU+format	+0.67 [0.27, 1.09]	+0.85 [0.15, 1.55]
Test 9B, concept task	IoU+format	+0.17 [−0.84, 1.17]	−0.40 [−1.70, 0.90]
Dev, all tasks	direct, step 50, 3 seeds	+0.06 (SD 0.08)	+0.24 (SD 0.15)
Dev, multi-view (n=59)	direct, step 50, 3 seeds	+4.45 (SD 0.63)	+7.35 (SD 1.96)
Dev, all tasks	task-balanced, threshold reward	+0.36 [−0.16, 0.89]	+0.94 [−0.07, 1.95]

6 Discussion and Conclusion

Sequential medical grounding is not simply single-image localization at scale: the main difficulty is deciding which cross-image relation matters before placing a box. Supervising that comparison with image-grounded rationales yields the best reported MedSG-Bench results at 4B, 9B and 27B, with the largest gains where the answer is defined by a cross-image comparison, while losses on appearance-matching tasks show that more reasoning is not always better. The intervention results indicate that the reasoning is not decorative, and GRPO with an overlap reward adds little once strong reasoning supervision is in place. Appendix E discusses limitations.

AI USE STATEMENT

In this work, we used generative AI tools to generate a synthetic dataset: the MedChain rationales were produced by FlemingVL-38B and Qwen3.5-397B-A17B as part of the research method (Section 3.1 and Appendix A), and a sample of 2,000 rationales was audited by a medical expert. We also used generative AI tools to assist with interpreting results, in drafting parts of the discussion of per-task findings (Appendix D); the authors checked every interpretation against the underlying experiment outputs. Additionally, we used generative AI tools for drafting and restructuring paper sections, editing for readability, suggesting paper structure, and formatting tables and LaTeX. We have reviewed all AI-assisted work and verified all numerical statements against the experiment artifacts. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study uses the publicly distributed MedSG-Bench research benchmark and does not introduce new patient recruitment or collect additional identifiers. The generated rationales can inherit clinical errors and biases from their teacher models, are conditioned on annotated boxes, and were reviewed only in a limited single-expert audit. Neither the rationales nor the grounding models are validated for autonomous diagnosis or clinical deployment. Users must respect the privacy, license and use restrictions of each underlying data source and obtain appropriate institutional approval for any prospective clinical study. Broader impacts are discussed in Appendix F.

REPRODUCIBILITY STATEMENT

Section 4, Section 5 and Appendix A report the data splits, evaluation protocol, optimizer settings, seeds, exact Qwen3.5 model identifiers, validation-selected checkpoint steps and the train–test overlap audit. Appendix B provides the rationale-generation prompts, and Appendix C records the additional baselines, GRPO configuration, reasoning-reliance controls and development-set studies. The supplementary material contains the training and evaluation scripts, the per-example predictions used for the paired bootstrap analyses, machine-readable overlap results and the checkpoint metadata needed to reproduce the reported analyses.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. 2025b. URL https://arxiv.org/abs/2502.13923.

Sule Bai, Mingxing Li, Yong Liu, Jing Tang, Haoji Zhang, Lei Sun, Xiangxiang Chu, and Yansong Tang. Univg-r1: Reasoning guided universal visual grounding with reinforcement learning. *arXiv preprint arXiv:2505.14231*, 2025c.

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. Learning to exploit temporal structure for biomedical vision-language processing. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15016–15027. IEEE, 2023.

Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, et al. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*, 2024.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pp. 1–21. Springer, 2022.

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*, pp. 7346–7370, 2024a.

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*, 2024b.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024c.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, et al. Chexagent: Towards a foundation model for chest x-ray interpretation. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*, 2024d.

David Dai, Peilin Chen, Chanakya Ekbote, and Paul Liang. Qoq-med: Building multimodal clinical foundation models with domain-aware grpo training. *Advances in Neural Information Processing Systems*, 38:37406–37453, 2026.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1749–1759. IEEE, 2021.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Wenxuan Huang, Bohan Jia, Shaosheng Cao, Zheyu Ye, Zhe Xu, Yao Hu, Shaohui Lin, et al. Vision-r1: Incentivizing reasoning capability in multimodal large language models. 2026:63794–63812, 2026.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 1760–1770. IEEE, 2021.

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.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

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.

You Li, Heyu Huang, Chi Chen, Kaiyu Huang, Chao Huang, Zonghao Guo, Zhiyuan Liu, Jinan Xu, Yuhua Li, Ruixuan Li, et al. Migician: Revealing the magic of free-form multi-image grounding in multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 9845–9867, 2025.

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

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

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*, pp. 337–347. Springer, 2025.

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.

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.

Yan Shu, Chi Liu, Robin Chen, Derek Li, and Bryan Dai. Fleming-vl: Towards universal medical visual reasoning with multimodal llms. *arXiv preprint arXiv:2511.00916*, 2025.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

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.

Huihui Xu, Yuanpeng Nie, Hualiang Wang, Ying Chen, Wei Li, Junzhi Ning, Lihao Liu, Hongqiu Wang, Lei Zhu, Jiyao Liu, et al. Medground-r1: Advancing medical image grounding via spatial-semantic rewarded group relative policy optimization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 391–401. Springer, 2025.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: Generalist foundation model for unified multimodal medical understanding and reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Yuan Yao, Tianyu Yu, Chongyi Wang, Junbo Cui, Bokai Xu, Hongji Zhu, Tianchi Cai, Fuwei Huang, Tianran Wang, Wenshuo Ma, et al. Minicpm-o 2.6: A gpt-4o level mllm for vision, speech, and multimodal live streaming on your phone. 2025.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. 2025:98891–98913, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026.

Jingkun Yue, Siqi Zhang, Zinan Jia, Huihuan Xu, Zongbo Han, Xiaohong Liu, and Guangyu Wang. Medsg-bench: A benchmark for medical image sequences grounding. *Advances in Neural Information Processing Systems*, 38, 2026.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 3: system demonstrations)*, pp. 400–410, 2024.

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.

Ke Zou, Yang Bai, Zhihao Chen, Yang Zhou, Yidi Chen, Kai Ren, Meng Wang, Xuedong Yuan, Xiaojing Shen, and Huazhu Fu. Medrg: Medical report grounding with multi-modal large language model. *arXiv e-prints*, pp. arXiv–2404, 2024.

Appendix

Table of Content

A MedChain-182K Dataset Construction Details **15**

A.1 Model Selection for Pseudo-CoT Generation 15

A.2 Stage 1: Pseudo-CoT Generation (FlemingVL-38B) 15

A.3 Stage 2: Multimodal CoT Refinement 15

A.4 Quality Verification 15

A.5 Train–Test Overlap Audit 16

A.6 Zero-Shot Evaluation Protocol 17

A.7 Coordinate-Convention Sensitivity of Qwen3.5 Zero-Shot Baselines 17

A.8 dCoT-SFT Training Hyperparameters 18

B MedChain Training Prompts **18**

B.1 Stage 1 Pseudo-CoT Generation Prompt (FlemingVL-38B) 18

B.2 Stage 2 Multimodal CoT Refinement Prompt 19

C Additional Results **20**

C.1 Task-Specific RL Baseline 20

C.2 Qwen2.5-VL-7B Continued Training 22

C.3 GRPO Configuration 22

C.4 Reasoning-Reliance Inference Controls 22

C.5 Task-wise Paired Bootstrap Comparisons 24

C.6 GRPO on the Full Benchmark: Task-wise Effects and Hypotheses 25

C.7 Controlled GRPO Development Studies 26

D Extended Discussion of Findings **27**

D.1 Design of the two-stage pipeline 27

D.2 Reading the task-level pattern 27

D.3 The answer-only baseline and MedSeq-Grounder 28

D.4 Model scale 28

D.5 Why GRPO adds little on top of dCoT-SFT 28

D.6 What box conditioning implies 29

D.7 Practical takeaways 29

E Limitations **29**

F Broader Impacts **29**

G Qualitative Case Studies **30**

H Extended Related Work **30**

A MEDCHAIN-182K DATASET CONSTRUCTION DETAILS

A.1 MODEL SELECTION FOR PSEUDO-COT GENERATION

To choose the Stage 1 teacher, we compared six candidate multimodal models (Table 4). Each candidate generated rationales for the same 100 sampled instances under six prompt variants, and we assessed the outputs for reasoning quality, consistency with the task and image sequence, and clinical plausibility. The candidates cover three groups: general-domain MLLMs, medical MLLMs, and the task-specific sequential grounding model.

Table 4: Candidate teachers for Stage 1 pseudo-CoT generation. All candidates were evaluated on the same 100 sampled instances with six prompt variants each. The selected teacher is highlighted.

Candidate model	Category	Parameters
General-domain MLLMs		
Qwen3-VL (Bai et al., 2025a)	General multimodal	30B
GPT-4o (ChatGPT API) (Achiam et al., 2023)	General multimodal	Not disclosed
Medical MLLMs		
Lingshu (Xu et al., 2026)	Medical multimodal	32B
QoQ-Med-VL (Dai et al., 2026)	Medical multimodal	32B
FlemingVL (selected) (Shu et al., 2025)	Medical multimodal	38B
Task-specific model		
MedSeq-Grounder (Yue et al., 2026)	Sequential medical grounding	7B

Selection. FlemingVL-38B (Shu et al., 2025) produced the most reliable, task-consistent and clinically faithful rationales across the ten imaging modalities of MedSG-Bench (Yue et al., 2026), and we therefore use it as the Stage 1 teacher. Stage 2 refinement uses a separate multimodal model, Qwen3.5-397B-A17B (Appendix A.3).

A.2 STAGE 1: PSEUDO-COT GENERATION (FLEMINGVL-38B)

Following model selection, we developed the final system prompt (see Appendix B.1) and deployed FlemingVL-38B Shu et al. (2025) via `vLLM`-based inference on NVIDIA A100 40 GB GPUs, consuming approximately **1,200 A100 GPU hours**. Due to the model's 8K context window limitation, samples with longer input sequences were isolated into error files and reprocessed in the subsequent stage.

A.3 STAGE 2: MULTIMODAL COT REFINEMENT

Qwen3.5-397B-A17B Team (2026) directly processed the original image sequence, question, task context, annotated box, and optional sanitized preliminary rationale to produce the final explanation (see Appendix B.2), executed on **NVIDIA H200 140 GB GPUs** and requiring approximately **1,280 H200 GPU hours**. Samples with missing pseudo-CoTs or incorrect coordinate associations were isolated and recursively reprocessed.

A.4 QUALITY VERIFICATION

One medical expert reviewed **2,000 rationales** for clinical plausibility, consistency with the task and image sequence, and correspondence to the intended target. Of these, 1,965 were accepted (98.25%; Wilson 95% CI 97.58–98.74%) and 35 contained minor-to-moderate CoT errors; the boxes were correct in all 35 flagged cases. Because only aggregate counts are available, error-category counts, item-level sampling verification, and inter-rater agreement cannot be reported. This is a single-rater audit, not multi-expert clinical validation.

The available audit record supports the following breakdown and no finer subdivision. We report missing metadata explicitly rather than reconstructing it after the review.

Table 6 summarizes the full pipeline statistics.

Table 5: Clinical rationale audit record. "Not recorded" denotes information absent from the retained audit summary.

Audit field	Retained result
Reviewer	One medical expert
Specialty/qualification	Not recorded
Clinical plausibility criterion	Assessed
Task/sequence consistency criterion	Assessed
Rationale–target correspondence	Assessed
Accepted	1,965 (98.25%)
Minor-to-moderate CoT error	35 (1.75%)
Incorrect box among flagged cases	0/35
Finer error subclasses/counts	Not recorded
Second rater / Cohen's κ	Not available

Table 6: MedChain-182K construction pipeline statistics.

Component	**Details**
Stage 1: Pseudo-CoT Generation	
Model	FlemingVL-38B
Hardware	NVIDIA A100 40 GB
Compute	∼1,200 A100 GPU hours
Context limit	8K tokens
Stage 2: CoT Refinement	
Model	Qwen3.5-397B-A17B (multimodal)
Hardware	NVIDIA H200 140 GB
Compute	∼1,280 H200 GPU hours
Dataset Statistics	
Filtered CoT pool	182,573 unique IDs
GRPO allocation	32,000 IDs (4,000/task)
Remaining SFT IDs	150,573
Released SFT artifact	152,396 rows
Tasks covered	8 clinical tasks
Modalities	10 modalities
Quality Verification	
Reviewed rationales	2,000
Accepted / flagged	1,965 / 35

A.5 Train–Test Overlap Audit

We hashed all referenced images in the 152,396-row SFT artifact and the 9,630-example test set using both encoded-file SHA-256 and decoded RGBA-pixel SHA-256. We additionally compared complete ordered image sequences and task-aware source series/cases. The task-aware normalization supersedes an earlier filename-only normalization that removed too much case information. Table 7 reports the verified counts.

Table 7: Detailed SFT/test overlap audit. Affected counts refer to examples, whereas shared counts refer to unique hashes, sequences, or source-series keys.

Overlap definition	Shared	Test
Encoded-file SHA-256	55	144
Decoded-pixel SHA-256	56	147
Complete encoded-hash sequence	0	0
Complete pixel-hash sequence	0	0
Task-aware source series/case	12	17
Series + normalized question	2	2
Series + concept	12	17
Series + target box	0	0
Series + concept + target box	0	0
Global target-box coordinates	100	106
Image/series union excluded	—	164
Leakage-clean paired subset	—	9,466

Table 8: dCoT-SFT training hyperparameters per model scale.

Hyperparameter	**4B**	**9B**	**27B**
Learning rate	5×10^{-6}	5×10^{-6}	5×10^{-6}
Per-device batch size	4	4	2
Gradient accumulation	2	2	4
Effective batch size	64	64	64
Epochs	2	2	6
Warmup ratio	3%	3%	1.5%
Weight decay	0.05	0.05	0.00
Best checkpoint step	2000	2500	4500
Best val. loss	0.2787	0.2633	0.3644

The 147 image-content matches comprise 135 referring, 10 multi-view, one object-tracking, and one cross-modal test example. No complete multi-image test input and no source-series-plus-target-box pair occurs in SFT. A global coordinate collision alone is not evidence of leakage. Conservatively excluding the union of all image-content and task-aware source-series matches leaves 9,466 paired examples; the dCoT-SFT minus answer-only differences remain $+3.605$ mIoU points (95% CI 3.167–4.043) and $+2.007$ Acc@0.5 points (1.310–2.704), with paired-bootstrap $p < 10^{-5}$ for both metrics.

A.6 Zero-Shot Evaluation Protocol

To ensure fair comparison across all baseline models Bai et al. (2025b); Zhu et al. (2025); Li et al. (2024); Sellergren et al. (2025), we adopt a unified zero-shot evaluation protocol on MedSG-Bench Yue et al. (2026). Each model receives the clinical grounding question with a model-specific output format instruction appended, requiring exactly one bounding box prediction with no chain-of-thought supervision.

Zero-Shot Evaluation Prompt (All Models)

```
{Clinical grounding question}

{Model-specific format suffix}

Example for Qwen3-VL:
Identify the bounding box of the region described by the following
expression: <|object_ref_start|> {target} <|object_ref_end|>
Output only the coordinates in integer pixel space. Format:
(x1,y1),(x2,y2). Strictly follow this format. No additional text
or explanation.

Example for generic models:
Identify the bounding box of the region described by the following
expression: {target}
Format: (x1,y1),(x2,y2). Output exactly one bounding box and
nothing else.
```

A.7 Coordinate-Convention Sensitivity of Qwen3.5 Zero-Shot Baselines

None of the 9,630 MedSG-Bench test questions specifies a coordinate space, and some Qwen vision-language models express boxes in a 0–1000 normalized frame, so near-zero Qwen3.5 scores could partly reflect a coordinate mismatch rather than a grounding failure. To test this, we score the answers of each Qwen3.5 zero-shot run under two conventions: reading the predicted box as pixel coordinates, and reading it as 0–1000 normalized coordinates rescaled by the target image width and height. The model is never asked for normalized coordinates, and the two rows of each run score an identical set of answers. Boxes are extracted with the benchmark evaluator's parser, which takes the first box in the output; answers without a parseable box score zero under both conventions. Table 9 covers two sources. The first is the Qwen3.5-4B thinking-mode run reported in Table 10, rescored. The second is a separate thinking-off rerun of Qwen3.5-4B, 9B and 27B on all 9,630 questions, in which the model is instructed only to output the box coordinates inside answer tags. The thinking-mode outputs behind the 9B and 27B rows of Table 10 were not retained, so they could not be rescored.

Table 9 reports overall scores. In the thinking-off rerun, normalized scoring raises overall mIoU from 1.55 to 8.21 at 4B, from 1.04 to 13.05 at 9B and from 2.64 to 13.30 at 27B, so the coordinate convention explains part of the zero-shot gap for these runs. It explains almost none of the gap for the thinking-mode 4B run (4.52 pixel versus 4.55 normalized mIoU). That run returns a parseable box for only 6,368 of the 9,630 questions, so its scores mainly reflect format failures, and its object-tracking score comes largely from restating the reference box given in the question. Scaling from 9B to 27B changes normalized-coordinate mIoU only slightly (13.05 to 13.30). Under either convention, the best zero-shot result (13.30 mIoU, Qwen3.5-27B) remains far below MedSeq-Grounder (72.55) and our dCoT-SFT models (74.96 to 77.24 mIoU), so the conclusion that sequential grounding requires task-specific training is unchanged.

Table 9: **Qwen3.5 zero-shot results under pixel and 0–1000 normalized coordinate scoring.** MedSG-Bench test set (9,630 questions), overall example-weighted scores in percent. Both columns of a row score the same set of answers; unparseable answers score zero. *Parsed*: questions with a parseable box.

		Pixel		Normalized		
Model	**Run**	**mIoU**	**Acc@0.5**	**mIoU**	**Acc@0.5**	**Parsed**
Qwen3.5-4B	thinking (Table 10)	4.52	2.96	4.55	2.14	6,368
Qwen3.5-4B	thinking off	1.55	0.86	8.21	2.83	8,139
Qwen3.5-9B	thinking off	1.04	0.61	13.05	6.23	9,630
Qwen3.5-27B	thinking off	2.64	1.64	13.30	7.64	9,466
Reference: MedSeq-Grounder (7B)		72.55	79.71	—	—	—

A.8 dCoT-SFT Training Hyperparameters

All models are fine-tuned using full-parameter supervised fine-tuning via LLaMA-Factory Zheng et al. (2024) with AdamW optimizer (β_1=0.9, β_2=0.999, ϵ=10^{-8}) and cosine learning rate schedule, seed 42, on 8 GPUs. Checkpoints are selected at minimum validation loss to prevent overfitting, as training loss continues to decrease beyond epoch 2 while validation loss rises monotonically. Table 8 summarizes the hyperparameters per model scale.

Exact model identifiers. The multimodal base checkpoints are `Qwen/Qwen3.5-4B`, `Qwen/Qwen3.5-9B`, and `Qwen/Qwen3.5-27B` Team (2026). The reported dCoT-SFT rows use validation-selected internal checkpoints `checkpoint-2000`, `checkpoint-2500`, and `checkpoint-4500`, respectively. The 4B and 9B GRPO rows use checkpoints selected on a held-out development set from runs of up to 3,000 steps. Stage 2 rationale generation uses the native-multimodal `Qwen/Qwen3.5-397B-A17B` checkpoint, not text-only Qwen3-235B.

B MedChain Training Prompts

B.1 Stage 1 Pseudo-CoT Generation Prompt (FlemingVL-38B)

The following prompt is provided to FlemingVL-38B Shu et al. (2025) during Stage 1. The model processes the sequential medical images directly and generates domain-aware pseudo-CoT descriptions, subsequently refined by multimodal Qwen3.5-397B-A17B Team (2026) in Stage 2.

Stage 1: Pseudo-CoT Generation Prompt (FlemingVL-38B)

You are a specialized clinical radiologist with expertise in interpreting medical image sequences across multiple modalities (CT, MRI, X-ray, ultrasound, etc.) and performing precise visual grounding.

You will be given:

- A sequence of medical images
- A clinical grounding question
- The correct bounding box of the target region (for verification only)

Your task is to generate a realistic, medically accurate Chain-of-Thought (CoT) that follows the reasoning process an experienced radiologist or clinician would use, written as smooth narrative prose rather than numbered steps.

The reasoning should naturally cover the following aspects, integrated into continuous paragraphs:

1. Identify the imaging modality and anatomical context.
Describe what type of imaging is used (CT, MRI, X-ray, ultrasound, etc.), the imaging plane if relevant (axial, sagittal, coronal), and the main body region and structures that are visible.

2. Interpret the clinical question.
Explain what the question is asking for: which organ, structure, or pathology is important, and what kind of difference, object, or region needs to be localized. Mention any spatial cues (e.g., left vs right, superior vs inferior) or clinical cues (lesion, edema, fracture, fluid, instrument, etc.).

3. Visually analyze the images.
Describe what you see in terms of medically meaningful patterns: intensity or density changes, shapes and borders, symmetry or asymmetry, presence of lesions, masses, edema, or other abnormalities. For sequence or multi-image tasks, explain how the images relate to each other and what changes across them.

4. Localize the target region and justify it.
Explain clearly which region best matches the question and why. Use anatomical landmarks and spatial relationships (e.g., "in the lower pole of the left kidney", "adjacent to the lateral ventricle"). Connect the visual appearance and location to the clinical description, and briefly discuss what the finding could represent (e.g., nodule, infarct, hemorrhage, edema, calculus, or instrument).

5. Conclude with the bounding box.
After finishing the reasoning, indicate that the described region corresponds to the ground-truth bounding box and provide the coordinates.

Write everything in natural, continuous sentences and paragraphs, without explicit step numbers, bullets, or headings. The reasoning should read like a concise radiology teaching explanation, not an instruction list.

Output format (STRICT):

```
<analysis>
[Natural, flowing medical reasoning.
No Step 1/2/3 and no bullet points.]
</analysis>
<answer> (x1, y1),(x2, y2) </answer>
```

Rules:

- Do NOT mention that the ground-truth bounding box was provided.
- Do NOT copy the answer directly without reasoning; the analysis must stand on its own as a realistic clinical interpretation.
- The explanation must clearly connect the image content and the clinical question to the localized region.
- The final bounding box in `<answer>` MUST exactly match the given ground-truth coordinates `({x1},{y1}),({x2},{y2})`.

B.2 Stage 2 Multimodal CoT Refinement Prompt

The following text accompanies the original image sequence for multimodal Qwen3.5-397B-A17B Team (2026) during Stage 2. The model directly inspects the images; the optional FlemingVL-38B block is supplementary rather than its only visual context. The annotated box is privileged offline-generation information.

Stage 2: Multimodal CoT Refinement Prompt

```
MedSG-Bench grounded explanation task
Case ID: {sample_id}
Task: {task}
{task_ctx}
Original medical image sequence: {images}

Clinical question:
{question}

Known correct bounding box:
({x1}, {y1}),({x2}, {y2})

Instruction:
The bounding box above is already correct. Explain why this
specific boxed region answers the question but behave as if you do
not know the answer yet. Focus on what is visible inside the box,
the surrounding anatomy, and the discriminative cue that makes this
region the correct target rather than nearby alternatives. Do not
mention that the box was given to you or that you are verifying it.

Do not search for coordinates or describe how you would estimate the
box position. Justify the box with concise, image-grounded clinical
reasoning.

Style requirements:
• Sound like a concise human clinician or radiology teacher with
self-reflection questions and corrections (one or two only).
• Prefer concrete visual evidence over generic medical textbook
wording.
• Keep it short, specific, and image-grounded.
• Do not mention that the box was given for supervision or
verification.
• End by repeating the correct box in <answer>.

{prior_block}

Return exactly:
<think>...</think>
<answer>({x1}, {y1}),({x2}, {y2})</answer>
```

C Additional Results

Table 12 reports performance of additional general-domain and medical-specialized MLLMs on MedSG-Bench Yue et al. (2026) under the zero-shot evaluation protocol described in Appendix A.6. These results complement Table 10 and further confirm the substantial sequential grounding gap across all model families, including recent models such as Qwen3-VL Bai et al. (2025a) and MedGemma Sellergren et al. (2025).

C.1 Task-Specific RL Baseline

We evaluated MedGRPO-7B-RL with direct box generation on all 9,630 test examples without parser failures. Table 13 shows that training for adjacent surgical video frames transfers poorly to longitudinal, cross-view, and cross-modal medical sequences.

Table 10: Full MedSG-Bench results. The fixed 4B dCoT-SFT checkpoint is in Table 2. Avg.: Average; IoU and acc@0.5 for all results are shown, all numbers are in percentages.

Model	Size	IDG		ICG						Avg.
		RDG	NRDG	MV	OT	VCG	VPG	CMG	RG	
General-purpose MLLMs										
Qwen2.5-VLBai et al. (2025b)	3B	0.59 0.30	1.62 1.30	7.12 3.90	21.32 16.80	6.98 0.80	27.36 3.40	10.02 1.65	12.99 6.82	10.94 4.20
Qwen2.5-VLBai et al. (2025b)	7B	0.88 0.30	1.25 0.00	8.48 3.73	22.41 17.80	4.22 1.00	28.87 5.70	16.29 4.45	12.58 6.21	12.31 4.90
Qwen2.5-VLBai et al. (2025b)	32B	2.69 1.40	3.48 1.20	7.35 2.61	19.12 13.40	6.53 1.30	26.92 7.10	12.59 4.90	18.71 11.67	12.47 5.71
Qwen2.5-VLBai et al. (2025b)	72B	4.37 2.60	3.46 0.80	7.22 2.78	13.11 7.70	10.33 3.50	26.45 6.30	16.32 7.00	20.19 14.10	13.35 6.12
MiniCPM-V-2_6Yao et al. (2024)	8B	1.36 0.00	1.50 0.00	15.82 5.20	24.03 18.50	9.90 2.10	28.65 12.20	12.72 3.30	12.44 3.64	13.24 5.27
MiniCPM-O-2_6Yao et al. (2025)	8B	1.69 0.10	1.63 0.00	12.11 2.43	15.25 9.60	9.88 1.70	22.96 9.20	9.53 2.35	8.82 2.02	10.12 3.23
mPLUG-Owl3Ye et al. (2025)	7B	2.12 0.00	2.55 0.00	15.64 3.64	15.62 4.40	6.80 0.80	30.42 3.60	17.06 4.80	11.92 5.47	13.22 3.19
Mantis-Idefics2Jiang et al. (2024)	8B	0.49 0.00	0.62 0.00	18.69 8.59	28.04 23.50	6.27 0.50	10.26 1.10	9.59 0.95	6.05 0.54	9.90 3.91
LLaVA-OneVisionLi et al. (2024)	7B	1.09 0.00	0.01 0.00	9.26 1.13	10.50 3.20	11.33 1.80	22.20 5.30	19.08 6.70	17.11 5.67	12.39 3.47
LLaVA-OneVisionLi et al. (2024)	72B	2.58 0.80	2.87 0.90	11.74 1.39	9.61 2.30	10.95 3.30	32.38 20.30	16.24 5.40	15.43 6.68	13.21 5.18
InternVL2Chen et al. (2024c)	8B	0.18 0.00	0.38 0.00	17.34 7.03	26.45 21.20	5.56 0.80	10.36 0.70	6.23 1.00	15.73 7.69	10.24 4.59
InternVL2Chen et al. (2024c)	76B	0.15 0.00	0.15 0.00	10.00 3.90	15.56 11.80	3.39 0.40	6.64 1.10	2.83 0.75	15.69 9.92	6.88 3.53
InternVL2.5Chen et al. (2024b)	8B	0.26 0.00	0.38 0.00	13.52 3.56	20.82 13.80	1.96 0.00	5.25 0.00	4.70 0.85	9.56 3.44	7.04 2.56
InternVL2.5Chen et al. (2024b)	78B	0.24 0.10	0.32 0.10	9.16 2.08	16.18 10.00	4.32 0.50	11.86 2.30	5.48 1.25	10.67 4.52	7.29 2.55
InternVL3Zhu et al. (2025)	8B	1.07 0.30	1.20 0.00	14.36 4.42	13.30 6.50	6.43 0.90	18.73 4.60	4.73 1.15	15.16 7.42	9.26 3.19
InternVL3Zhu et al. (2025)	14B	0.66 0.00	0.71 0.00	13.24 5.31	19.77 13.00	8.60 2.10	13.17 2.40	10.87 3.70	14.57 7.76	10.53 4.41
InternVL3Zhu et al. (2025)	38B	0.98 0.10	1.76 0.20	12.99 4.79	19.27 13.60	7.63 2.10	17.76 2.90	6.47 1.75	16.59 10.05	10.37 4.44
InternVL3Zhu et al. (2025)	78B	0.20 0.00	0.53 0.00	6.35 2.43	13.03 8.00	3.57 0.90	11.81 2.50	3.34 0.85	12.76 8.10	6.44 2.90
MigicianLi et al. (2025)	7B	15.26 7.80	14.49 6.10	18.16 7.84	21.38 14.90	14.23 7.20	28.87 13.70	21.41 12.15	25.30 18.02	20.29 11.39
Medical-domain specialized MLLMs										
HuatuoGPT-VisionChen et al. (2024a)	7B	1.35 0.00	1.84 0.20	10.42 2.78	14.57 9.20	7.99 0.80	15.52 2.30	9.46 2.15	9.60 1.82	8.97 2.36
HuatuoGPT-VisionChen et al. (2024a)	34B	1.44 0.00	2.15 0.00	9.41 1.65	13.25 8.30	6.43 0.70	14.53 1.40	10.60 2.60	8.60 1.75	8.57 2.09
MedSeq-Grounder	7B	83.29 93.20	83.72 94.10	55.03 60.19	62.10 67.20	74.11 82.60	85.25 98.80	78.77 82.75	60.43 65.59	72.55 79.71
Ours-dCoT+GRPO (4B)	4B	**83.82** 94.50	83.21 92.70	54.80 59.93	**67.90** **76.20**	66.09 72.10	**89.43** **99.80**	**83.59** **87.55**	**70.72** **80.30**	**75.32** **82.89**
Ours-CoT-SFT (9B)	9B	**84.00** **93.10**	83.02 91.60	**55.87** **60.98**	**67.46** **75.00**	**74.64** **83.20**	**91.22** **99.90**	**84.16** **87.40**	**70.77** **78.95**	**76.60** **83.55**
Ours-dCoT+GRPO (9B)	9B	**84.16** 93.20	83.12 91.40	**56.98** **62.02**	**67.65** **74.90**	**74.80** **82.80**	**91.22** **99.80**	**84.82** **88.25**	**70.53** **79.22**	**76.90** **83.82**
Ours-CoT-SFT (27B)	27B	81.87 93.00	80.97 91.60	**60.14** **66.72**	**69.53** **78.30**	**74.64** **83.30**	**91.87** **99.90**	**84.24** **88.15**	**72.51** **81.98**	**77.24** **85.20**

Table 11: Ablation of training stages for MEDCHAIN-R1 on MedSG-Bench across model scales. IoU and `acc@0.5` are reported, all numbers in percentages.

Model	Size	IDG		ICG						Avg.
		RDG	NRDG	MV	OT	VCG	VPG	CMG	RG	
Qwen3.5 — 4B										
Zero-shot	4B	0.00 0.00	0.00 0.00	8.27 4.88	26.66 22.60	1.50 0.00	2.65 0.10	0.35 0.00	1.69 0.13	4.52 2.96
+ dCoT-SFT+GRPO	4B	**83.82** 94.50	83.21 92.70	54.80 59.93	**67.90** 76.20	66.09 72.10	**89.43** 99.80	**83.59** 87.55	**70.72** 80.30	**75.32** 82.89
Qwen3.5 — 9B										
Zero-shot	9B	0.00 0.00	0.00 0.00	8.57 4.97	28.30 24.50	0.52 0.00	1.62 0.20	0.25 0.00	1.26 0.13	4.43 3.18
+ CoT-SFT	9B	**84.00** 93.10	83.02 91.60	**55.87** 60.98	**67.46** 75.00	**74.64** 83.20	**91.22** 99.90	**84.16** 87.40	**70.77** 78.95	**76.60** 83.55
+ dCoT-SFT+GRPO	9B	**84.16** 93.20	83.12 91.40	**56.98** 62.02	**67.65** 74.90	**74.80** 82.80	**91.22** 99.80	**84.82** 88.25	**70.53** 79.22	**76.90** 83.82
Qwen3.5 — 27B										
Zero-shot	27B	0.00 0.00	0.00 0.00	8.48 4.88	28.39 24.50	0.25 0.00	2.50 0.20	0.25 0.00	0.82 0.00	4.42 3.15
+ CoT-SFT	27B	81.87 93.00	80.97 91.60	**60.14** 66.72	**69.53** 78.30	**74.64** 83.30	**91.87** 99.90	**84.24** 88.15	**72.51** 81.98	**77.24** 85.20

Bold = best in column per scale. Top = IoU, bottom = acc@0.5. Table 2 separates the verified dCoT-SFT and GRPO checkpoints.

C.2 QWEN2.5-VL-7B CONTINUED TRAINING

Full-parameter CoT-SFT continued from the published MedSeq-Grounder checkpoint reaches 75.71 mIoU and 82.34 Acc@0.5, compared with the published checkpoint's 72.55/79.71. This is useful continued-training evidence, but not a matched answer-only-versus-CoT experiment: the initialization is already MedSeq-Grounder-trained and no equally trained answer-only continuation is available.

C.3 GRPO CONFIGURATION

The GRPO run uses 32,000 examples (eight tasks $\times$ 4,000; seed 42), eight rollouts per prompt, one policy iteration, equal implicitly summed IoU and binary-format rewards, KL coefficient $\beta = 0.04$, clipping $\epsilon = 0.20$, initial learning rate 10^{-6} with linear decay, AdamW, per-device batch size 8, gradient accumulation 2, bfloat16, and a 1,200-token completion limit. The effective sampling temperature is 1.0.

C.4 REASONING-RELIANCE INFERENCE CONTROLS

We evaluated two inference-time interventions on the same 4B dCoT-SFT validation-selected checkpoint-2000 used for the 74.955 mIoU / 82.368 Acc@0.5 baseline. Both controls use the identical 9,630-example test set and deterministic decoding. In the *no-think* condition, the assistant response is prefixed with `<answer>` so the model must immediately generate a box. In the *corrupted-reasoning* condition, each example is deterministically assigned a rationale from a different task, followed by an `<answer>` prefix. There were no inference failures.

Assistant prefilling occasionally removes only the outer parentheses from an otherwise unambiguous coordinate string (for example, `<answer>90,142),(106,161</answer>`). We therefore apply one delimiter-tolerant answer-tag parser to the baseline and both interventions; this leaves the normal baseline score unchanged and prevents valid coordinates from being mapped to the zero box.

Table 12: Performance of additional MLLMs on MedSG-Bench. IoU and acc@0.5 are reported, all numbers in percentages.

Model	Size	IDG		ICG						Avg.
		RDG	NRDG	MV	OT	VCG	VPG	CMG	RG	
General-domain MLLMs										
Qwen3-VLBai et al. (2025a)	4B	0.01 0.00	0.00 0.00	7.27 3.83	25.08 21.30	0.62 0.00	2.67 0.00	0.20 0.00	0.94 0.00	4.00 2.67
Qwen3-VLBai et al. (2025a)	8B	0.04 0.00	0.00 0.00	7.43 3.83	24.73 20.60	0.74 0.00	1.87 0.00	0.46 0.05	1.12 0.00	4.00 2.61
Qwen3-VLBai et al. (2025a)	30B	0.01 0.00	0.00 0.00	3.12 1.66	0.65 0.40	0.82 0.00	2.50 0.00	0.10 0.00	1.05 0.00	0.97 0.24
GemmaTeam et al. (2024)	4B	0.03 0.00	0.00 0.00	0.24 0.09	0.95 0.80	0.79 0.00	3.12 0.00	0.11 0.05	1.09 0.00	0.73 0.10
GemmaTeam et al. (2024)	26B	0.01 0.00	0.00 0.00	0.59 0.26	2.83 2.70	0.17 0.00	1.26 0.00	0.02 0.00	0.59 0.00	0.61 0.31
Medical-domain specialized MLLMs										
Med-VLM-R1Pan et al. (2025)	2B	0.01 0.00	0.01 0.00	4.16 2.53	8.98 7.00	0.06 0.00	0.68 0.00	0.05 0.00	0.35 0.00	1.57 1.03
MedGemmaSellergren et al. (2025)	1.5B	0.03 0.00	0.01 0.00	8.53 4.97	28.46 24.40	0.91 0.00	2.08 0.10	0.46 0.00	1.01 0.00	4.54 3.14
MedGemmaSellergren et al. (2025)	4B	0.58 0.00	0.26 0.00	5.67 2.44	18.86 11.70	9.42 1.30	7.06 1.30	3.75 0.95	8.49 1.28	6.52 2.17
MedGemmaSellergren et al. (2025)	27B	0.01 0.00	0.06 0.00	5.29 2.26	22.26 16.10	2.27 0.00	3.11 0.30	2.92 0.35	5.35 0.61	4.94 2.14
Fleming VLShu et al. (2025)	8B	0.05 0.00	0.03 0.00	6.50 3.57	22.57 18.20	1.48 0.10	2.92 0.10	0.89 0.10	0.00 0.00	3.77 2.36
Lingshu Xu et al. (2026)	7B	0.20 0.00	0.03 0.00	4.94 1.83	19.53 16.00	3.47 0.10	18.30 1.00	3.48 0.10	4.22 1.69	6.27 2.27
Lingshu Xu et al. (2026)	32B	0.04 0.00	0.02 0.00	4.54 1.74	17.54 13.20	1.34 0.20	0.03 0.00	0.59 0.00	0.00 0.00	2.63 1.60
QoQ-Med3 Dai et al. (2026)	8B	0.04 0.00	0.00 0.00	7.35 3.75	24.92 21.00	0.66 0.00	1.64 0.00	0.50 0.10	1.12 0.00	3.98 2.65

Top = IoU, bottom = acc@0.5.

Table 13: MedGRPO-7B-RL transfer to MedSG-Bench; mIoU/Acc@0.5 in percent.

Task	Score	Task	Score
RDG	0.09/0.00	NRDG	0.06/0.00
Multi-view	3.68/1.48	Object Tracking	17.63/15.00
Concept	4.57/0.40	Patch	18.44/1.20
Crossmodal	0.78/0.15	Referring	9.31/3.37
		Overall	6.27/2.45

Table 14: Reasoning-reliance controls on the fixed 4B dCoT-SFT checkpoint; all values are percentages.

Inference condition	mIoU	Acc@0.5
Normal dCoT inference	74.955	82.368
Forced no-think	72.025	78.660
Corrupted cross-task reasoning	62.773	68.930

Table 15: Paired intervention-minus-normal differences from 10,000 example-level bootstrap resamples. Differences and confidence intervals are percentage points.

Intervention	Δ mIoU [95% CI]	Δ Acc@0.5 [95% CI]
No-think	-2.930 $[-3.293, -2.580]$	-3.707 $[-4.309, -3.126]$
Corrupted	-12.182 $[-12.745, -11.603]$	-13.437 $[-14.216, -12.648]$

Table 16: Task-wise reasoning-control results on the fixed 4B dCoT-SFT checkpoint. Each cell reports mIoU/Acc@0.5 in percent.

Task	Normal dCoT	Forced no-think	Corrupted reasoning
Registered Difference	82.86/93.30	82.15/92.20	65.19/73.30
Non-Registered Difference	81.30/91.20	80.94/90.80	80.16/89.60
Multi-view	54.01/58.97	49.72/53.57	39.55/41.55
Object Tracking	64.77/71.00	55.26/58.90	46.74/50.60
Concept	70.64/78.80	69.40/76.40	65.12/72.40
Patch	90.21/99.70	87.55/99.00	86.03/96.80
Crossmodal	83.39/87.00	81.44/84.85	65.78/70.00
Referring	69.67/79.28	66.36/73.55	56.88/63.02
Overall	74.96/82.37	72.02/78.66	62.77/68.93

All paired bootstrap and sign-flip tests reach $p < 10^{-4}$. Removing generated reasoning causes a moderate but significant decline, while an unrelated reasoning prefix causes a substantially larger decline. This supports sensitivity to reasoning context, not the stronger claim that the displayed rationale is a faithful causal trace. Corrupted prefilling also constitutes a deliberate distribution shift. These controls use the SFT checkpoint rather than the GRPO checkpoint because the retained local GRPO weights are incomplete; they must therefore be compared with the 74.955/82.368 SFT baseline, not the 75.321/82.887 GRPO result.

C.5 Task-wise Paired Bootstrap Comparisons

Tables 17 and 18 compare the fixed 4B dCoT-SFT checkpoint-2000 and the GRPO checkpoint with the fixed 4B answer-only SFT checkpoint on each task of the 9,630-example test set. Predictions are paired by example ID, and each task uses 100,000 paired example-level bootstrap resamples. Differences are the comparison model minus answer-only SFT, in percentage points. Each task contains 1,000 examples except multi-view (1,148), cross-modal (2,000) and referring (1,482); the overall scores in Tables 1 and 2 are example-weighted means over these tasks. With 16 tests per table, a Bonferroni-adjusted threshold is $0.05/16 \approx 0.003$. Reported $p < 10^{-5}$ indicates that no resample reached the null.

The dCoT-SFT difference on concept grounding is not significant ($p = 0.063$ for mIoU, $p = 0.78$ for Acc@0.5). On patch grounding, mIoU decreases while Acc@0.5 is unchanged (99.7 for both models), so the loss reflects box tightness rather than missed targets. On multi-view and referring grounding, the Acc@0.5 decreases (-4.27 and -4.18) are larger than the corresponding mIoU decreases. GRPO remains below answer-only SFT on concept, multi-view, patch and referring grounding.

Table 17: **Task-wise paired comparison: dCoT-SFT versus answer-only SFT (4B).** Paired example-level bootstrap on the 9,630 MedSG-Bench test questions (100,000 resamples per task). Cells show mIoU (top) and Acc@0.5 (bottom) in percent; Δ rows give the comparison minus answer-only SFT in percentage points, computed from unrounded scores, with 95% CIs and p-values. Significant gains ($p < 0.05$) are shaded green and significant losses red. The Overall column repeats the paired analysis of Section 5.2.

Qwen3.5-4B *N*	**RDG** 1,000	**NRDG** 1,000	**MV** 1,148	**OT** 1,000	**VCG** 1,000	**VPG** 1,000	**CMG** 2,000	**RG** 1,482	**Overall** 9,630
Answer-only SFT	66.76 81.00	66.64 81.40	56.21 63.24	58.27 66.40	69.03 78.40	92.66 99.70	80.83 85.10	72.85 83.47	71.39 80.31
dCoT-SFT (ckpt. 2000)	82.86 93.30	81.30 91.20	54.01 58.97	64.77 71.00	70.64 78.80	90.21 99.70	83.39 87.00	69.67 79.28	74.96 82.37
Δ mIoU	+16.10	+14.67	−2.20	+6.50	+1.61	−2.45	+2.56	−3.18	+3.56
95% CI	[14.94, 17.28]	[13.54, 15.80]	[−3.53, −0.87]	[4.72, 8.29]	[−0.10, 3.30]	[−2.81, −2.11]	[1.81, 3.31]	[−4.25, −2.13]	[3.13, 4.00]
p	$< 10^{-5}$	$< 10^{-5}$	0.0010	$< 10^{-5}$	0.0634	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$
Δ Acc@0.5	+12.30	+9.80	−4.27	+4.60	+0.40	+0.00	+1.90	−4.18	+2.06
95% CI	[10.20, 14.40]	[7.90, 11.80]	[−6.62, −1.92]	[1.80, 7.40]	[−2.00, 2.80]	[−0.40, 0.40]	[0.60, 3.20]	[−6.07, −2.36]	[1.36, 2.75]
p	$< 10^{-5}$	$< 10^{-5}$	0.0004	0.0013	0.7772	1.0000	0.0041	$< 10^{-5}$	$< 10^{-5}$

Table 19 applies the same paired analysis to GRPO at 9B, comparing the GRPO checkpoint with its 9B dCoT-SFT starting point (checkpoint-2500). GRPO significantly improves multi-view (+1.11

Table 18: **Task-wise paired comparison: GRPO versus answer-only SFT (4B).** Paired example-level bootstrap on the 9,630 MedSG-Bench test questions (100,000 resamples per task). Cells show mIoU (top) and Acc@0.5 (bottom) in percent; Δ rows give the comparison minus answer-only SFT in percentage points, computed from unrounded scores, with 95% CIs and p-values. Significant gains ($p < 0.05$) are shaded green and significant losses red. An overall paired interval was not computed for this comparison (—).

Qwen3.5-4B N	**RDG** 1,000	**NRDG** 1,000	**MV** 1,148	**OT** 1,000	**VCG** 1,000	**VPG** 1,000	**CMG** 2,000	**RG** 1,482	**Overall** 9,630
Answer-only SFT	66.76 81.00	66.64 81.40	56.21 63.24	58.27 66.40	69.03 78.40	92.66 99.70	80.83 85.10	72.85 83.47	71.39 80.31
GRPO	83.82 94.50	83.21 92.70	54.80 59.93	67.90 76.20	66.09 72.10	89.43 99.80	83.59 87.55	70.72 80.30	75.32 82.89
Δ mIoU	+17.05	+16.57	−1.41	+9.63	−2.93	−3.24	+2.76	−2.12	—
95% CI	[15.88, 18.25]	[15.39, 17.77]	[−2.71, −0.10]	[7.93, 11.37]	[−4.65, −1.24]	[−3.57, −2.91]	[2.00, 3.51]	[−3.15, −1.11]	—
p	$< 10^{-5}$	$< 10^{-5}$	0.0341	$< 10^{-5}$	0.0007	$< 10^{-5}$	$< 10^{-5}$	$< 10^{-5}$	—
Δ Acc@0.5	+13.50	+11.30	−3.31	+9.80	−6.30	+0.10	+2.45	−3.17	—
95% CI	[11.40, 15.70]	[9.30, 13.40]	[−5.75, −0.87]	[7.10, 12.50]	[−8.80, −3.90]	[−0.20, 0.50]	[1.20, 3.70]	[−5.06, −1.35]	—
p	$< 10^{-5}$	$< 10^{-5}$	0.0098	$< 10^{-5}$	$< 10^{-5}$	0.7615	0.0002	0.0009	—

mIoU, $p = 0.0025$) and cross-modal grounding (+0.67 mIoU, $p = 0.0018$), and no task shows a significant decrease. Concept grounding is unchanged (+0.17 mIoU, $p = 0.74$), in contrast to the 4B decrease. The overall change is +0.30 mIoU and +0.27 Acc@0.5 (example-weighted means; no paired overall interval was computed).

Table 19: **Task-wise paired comparison: GRPO versus dCoT-SFT (9B).** Paired example-level bootstrap on the 9,630 MedSG-Bench test questions (100,000 resamples per task). Cells show mIoU (top) and Acc@0.5 (bottom) in percent; Δ rows give GRPO minus the dCoT-SFT checkpoint-2500 in percentage points, computed from unrounded scores, with 95% CIs and p-values. Significant gains ($p < 0.05$) are shaded green and significant losses red. The Overall Δ is a point estimate from example-weighted means; a paired overall interval was not computed (—).

Qwen3.5-9B N	**RDG** 1,000	**NRDG** 1,000	**MV** 1,148	**OT** 1,000	**VCG** 1,000	**VPG** 1,000	**CMG** 2,000	**RG** 1,482	**Overall** 9,630
dCoT-SFT (ckpt. 2500)	84.00 93.10	83.02 91.60	55.87 60.98	67.46 75.00	74.64 83.20	91.22 99.90	84.16 87.40	70.77 78.95	76.60 83.55
GRPO	84.16 93.20	83.12 91.40	56.98 62.02	67.65 74.90	74.80 82.80	91.22 99.80	84.82 88.25	70.53 79.22	76.90 83.82
Δ mIoU	+0.16	+0.10	+1.11	+0.19	+0.17	−0.00	+0.67	−0.24	+0.30
95% CI	[−0.21, 0.51]	[−0.23, 0.40]	[0.39, 1.83]	[−0.49, 0.89]	[−0.84, 1.17]	[−0.07, 0.07]	[0.27, 1.09]	[−0.75, 0.26]	—
p	0.3642	0.5353	0.0025	0.5806	0.7443	0.9362	0.0018	0.3490	—
Δ Acc@0.5	+0.10	−0.20	+1.05	−0.10	−0.40	−0.10	+0.85	+0.27	+0.27
95% CI	[−0.70, 0.90]	[−0.70, 0.30]	[−0.44, 2.53]	[−1.60, 1.50]	[−1.70, 0.90]	[−0.30, 0.00]	[0.15, 1.55]	[−0.81, 1.35]	—
p	0.8960	0.5348	0.1907	0.9505	0.6053	0.6356	0.0185	0.6592	—

C.6 GRPO ON THE FULL BENCHMARK: TASK-WISE EFFECTS AND HYPOTHESES

Figure 3 compares the fixed 4B dCoT-SFT checkpoint with the GRPO checkpoint on every task. The paired aggregate gains are +0.367 mIoU points (95% CI 0.049–0.686, $p = 0.0247$) and +0.519 Acc@0.5 points (−0.042–1.080, $p = 0.0746$). We therefore claim a small mIoU improvement, not a significant Acc@0.5 gain, and not a general failure of GRPO.

Reward-variance hypothesis. dCoT-SFT can produce low-diversity rollout groups, which reduces within-group reward variance and hence the GRPO advantage signal (Shao et al., 2024; Yu et al., 2026). The sampling diagnostic in Appendix C.7 shows that diversity is not absent on most training prompts, so this is at most a partial explanation.

Cross-task imbalance hypothesis. Starting mIoU spans 54 to 90 points across the eight tasks. Such heterogeneity motivates task-balanced sampling, but score differences alone do not show that easy tasks dominate the gradients. Our task-balanced follow-up did not produce a significant aggregate gain.

Semantic drift. The IoU+format reward improves most localization-style tasks but lowers concept grounding by 4.55 mIoU and 6.70 Acc@0.5. Spatial overlap rewards do not directly preserve the

clinical identification that concept grounding requires (Zou et al., 2024; Boecking et al., 2022; Xu et al., 2025). These observations motivate hypotheses about exploration, task imbalance and semantic preservation but do not isolate their causal contributions.

C.7 Controlled GRPO Development Studies

Scope and preprocessing control. These experiments use a 1,385-example development set, not the 9,630-example benchmark, and initialize from the 4B dCoT-SFT checkpoint-2000. An early GRPO export saved a processor with a reduced image-resolution limit. A zero-learning-rate save/reload control reproduced an apparent regression with that processor, but exactly reproduced all 1,385 SFT predictions with the original processor. The early 3–5-point regressions therefore cannot be attributed to learned policy degradation. Native-resolution retraining uses the SFT bounds of 65,536 minimum and 16,777,216 maximum pixels. Historical batch-8 and later batch-1 results are kept separate.

Fixed-checkpoint replication across seeds. Native-resolution direct GRPO was evaluated at step 50, fixed before evaluating the two additional seeds. Under the historical batch-8 protocol, the three-seed mean improves over SFT by only $+0.064$ mIoU and $+0.241$ Acc@0.5 points (unrounded differences).

Table 20: Fixed step-50 direct GRPO on the development set, historical batch-8 evaluation. Values are percentages; SD is the sample standard deviation across training seeds.

Run	mIoU	Acc@0.5
CoT-SFT	78.038	86.065
GRPO, seed 42	78.189	86.426
GRPO, seed 43	78.088	86.137
GRPO, seed 44	78.028	86.354
Mean $\pm$ SD	78.101 ± 0.081	86.306 ± 0.150

Table 21: Positive development point estimates with original prompts. All rows use identical examples, the SFT processor, and batch-1 evaluation. Step 100 is a secondary checkpoint; step 200 was the preprocessing ablation's primary endpoint.

Model	mIoU	Acc@0.5	Δ mIoU	Δ Acc
CoT-SFT	77.877	85.560	—	—
No-DIoU, step 100	78.235	86.498	+0.357	+0.939
No-DIoU, step 200	78.202	86.209	+0.325	+0.650

Multi-view is the clearest positive task-level pattern: mean mIoU increases from 53.589 to 58.038 ± 0.625 ($+4.449$ points), and mean Acc@0.5 from 57.627 to 64.972 ± 1.957 ($+7.345$ points). This exploratory result covers only 59 development examples; other tasks do not uniformly improve. It does not establish an overall 4–5-point gain or eliminate the historical batch-8 caveat.

Task-balanced, threshold-aware GRPO. A follow-up uses 15,000 RL examples, 3,750 each from Concept, Multi-view, Object Tracking, and Referring. The reward is IoU plus 0.2 times a sigmoid threshold reward centered at IoU 0.5 (temperature 0.05), plus 0.01 times the format reward; the DIoU coefficient is zero. Settings are eight rollouts, two policy iterations, centered within-group advantages, task-balanced sampling, no replay, all-token policy loss, frozen vision encoder, learning rate 5×10^{-7}, cosine decay over 200 updates, warmup ratio 0.05, $\beta = 0.04$, $\epsilon = 0.2$, and seed 42. Eight GPUs use per-device batch 2, accumulation 4, and a 768-token completion limit. Evaluation uses batch 1.

For step 100 versus SFT, 100,000 paired example-bootstrap resamples yield $+0.357$ mIoU points (95% CI $[-0.164, 0.886]$, $p = 0.182$) and $+0.939$ Acc@0.5 points ($[-0.072, 1.949]$, $p = 0.085$). Neither aggregate gain is significant at 0.05. These intervals quantify example sampling, not repeated-training uncertainty.

Prompt sensitivity and a negative control. Removing literal image-marker text, while retaining image inputs, changes SFT to 78.212/86.570 and step-200 GRPO to 78.196/86.282. The step-200 advantage therefore does not persist under this condition. A cleaned-prompt step-100 comparison is not yet available. A separate 16-update mixed-success-weighting pilot on 507 fixed development examples also does not beat SFT (75.434/84.221 versus 75.667/85.010). These qualifications prevent selecting only favorable checkpoints or evaluation conventions.

Sampling headroom, not deployable accuracy. On 128 training-only diagnostic prompts (32 per hard task), eight stochastic completions yield sampled pass@1 67.676% and pass@8 84.375%, with success defined as IoU ≥ 0.5. Multi-view yields 46.484%/62.500%, and Object Tracking 67.969%/93.750%. Sampled pass@1 averages the eight completions, rather than greedy decoding. These are neither held-out benchmark results nor a deployable oracle selector.

Inference repeatability. On 54 outcome-selected diagnostic examples (38 changed cases and 16 unchanged controls), repeated greedy inference with one fixed GRPO checkpoint yields zero text, box, or Acc@0.5 changes, both on the same node and across two nodes. This supports repeatability only under those settings, not clinical correctness or training-seed reproducibility.

Full-pass experiment status. The later 30,000-example, eight-task no-DIoU run is not reported as a completed result. No performance claim is made before its training and benchmark evaluation finish.

D Extended Discussion of Findings

This section collects clarifications and analyses that support the main text but did not fit within the page limit. All numbers are taken from results reported elsewhere in the paper; no new experiments are introduced here.

D.1 Design of the two-stage pipeline

MedChain separates *seeing* from *reasoning*. In Stage 1, FlemingVL-38B, a medical vision-language model, describes the modality, anatomy and visible findings of each image. These descriptions are medically accurate but read like captions: each image is described on its own, and the relation between images, which is what the grounding question asks about, is rarely stated. In Stage 2, Qwen3.5-397B-A17B receives the original image sequence together with the question, task context, annotated box and the Stage 1 description, and writes the final rationale. Because the refiner inspects the images directly, it can state the cross-image comparison that decides the answer, for example a density present in one scan and absent in its prior, or a structure that keeps its position relative to the mediastinum across slices, and it can add brief self-checks such as ruling out a windowing artifact (Figure 5). Every statement in the final rationale therefore remains tied to visible evidence, while the reasoning itself comes from the stage that can see the whole sequence.

D.2 Reading the task-level pattern

Table 22 brings together, for each task, the effect of reasoning supervision (dCoT-SFT minus answer-only SFT), the two inference-time interventions on the dCoT-SFT checkpoint, and the effect of GRPO on the full benchmark. All values are mIoU differences in percentage points on the 9,630-example test set, computed from the rounded entries in Tables 2 and 16 and Figure 3.

Table 22: Per-task synthesis of the 4B analyses (mIoU, percentage points). *Training effect*: dCoT-SFT minus answer-only SFT. *No-think* and *Cross-task*: change when the dCoT-SFT checkpoint is forced to answer immediately or is given a rationale from another task. *GRPO*: GRPO minus dCoT-SFT.

Task	Training effect	No-think	Cross-task	GRPO
Registered difference (RDG)	$+16.10$	-0.71	-17.67	$+0.96$
Non-registered difference (NRDG)	$+14.66$	-0.36	-1.14	$+1.91$
Multi-view (MV)	-2.20	-4.29	-14.46	$+0.79$
Object tracking (OT)	$+6.50$	-9.51	-18.03	$+3.13$
Concept (VCG)	$+1.61$	-1.24	-5.52	-4.55
Patch (VPG)	-2.45	-2.66	-4.18	-0.78
Cross-modal (CMG)	$+2.56$	-1.95	-17.61	$+0.20$
Referring (RG)	-3.18	-3.31	-12.79	$+1.05$
Overall	$+3.56$	-2.93	-12.18	$+0.37$

Several observations follow from this view. We state them as hypotheses, since each rests on single training runs.

Object tracking depends on reasoning at every stage. It gains from reasoning supervision, is the task most harmed by removing the reasoning (-9.51) or replacing it (-18.03), and gains most from GRPO ($+3.13$). Tracking a structure across slices requires explicitly carrying anatomical context from one image to the next, which is what the rationales describe.

Training benefit and inference-time reliance can dissociate. Non-registered difference grounding shows one of the largest training effects ($+14.66$) but is almost unaffected by removing (-0.36) or replacing (-1.14) the rationale. Registered difference grounding, by contrast, is strongly affected by a mismatched rationale (-17.67). One reading is that for non-registered pairs, reasoning supervision acts mainly as a training signal that shapes the model's visual representation, rather than as a computation the model performs at inference time. A second, not exclusive, reading is that the answer-only baseline is particularly weak on difference grounding (Appendix D.3), which inflates the measured training effect. Distinguishing these explanations requires matched multi-seed runs.

Some tasks rely on reasoning that does not help them. Multi-view and referring grounding are both hurt by reasoning supervision (-2.20 and -3.18), yet the trained model depends heavily on its rationale for these tasks (cross-task replacement costs -14.46 and -12.79). The model uses the reasoning, but the reasoning it learned for these tasks appears less useful than direct localization. This is consistent with the rationales being less informative for these tasks, for example because correspondences between mammographic views or between a phrase and a small region are harder to verbalize than a change between two scans. Multi-view is also the task where development-set GRPO produced its only consistent gain (Appendix C.7), which suggests that reward optimization can partly correct reasoning that supervised rationales teach imperfectly.

Concept grounding is where the overlap reward fails. Concept grounding gains little from reasoning supervision, depends only moderately on its rationale, and is the one task that GRPO clearly harms (-4.55). This task requires recognizing the same lesion or structure across very different images; an IoU reward can be increased by boxes that overlap the target without identifying it, which is the failure mode discussed in Section 5.3.

D.3 The answer-only baseline and MedSeq-Grounder

The answer-only 4B model scores 66.76 and 66.64 mIoU on registered and non-registered difference grounding, well below MedSeq-Grounder (83.29 and 83.72), which is also trained without rationales. The two answer-only models differ in backbone (Qwen3.5-4B versus Qwen2.5-VL-7B), training data (the MedChain SFT split versus MedSG-188K) and training schedule, so their gap is not attributable to any single factor. The practical consequence is that the $+14.7$ to $+16.1$ mIoU training effect on difference grounding in Table 2 measures the benefit of reasoning supervision *relative to our answer-only run*, and may overstate the benefit relative to a stronger answer-only model. Two results bound this concern. First, the 4B dCoT-SFT model comes within 2.5 mIoU of MedSeq-Grounder on difference grounding despite the smaller backbone, and exceeds it on four of the eight tasks and on average. Second, continuing dCoT-SFT from the MedSeq-Grounder checkpoint itself raises it from 72.55 to 75.71 mIoU (Appendix C.2), so reasoning supervision still helps when the starting point is the strongest answer-only model.

D.4 Model scale

Average performance increases monotonically with model size (74.96, 76.60 and 77.24 mIoU at 4B, 9B and 27B). The per-task picture is less uniform: from 9B to 27B, consistency tasks improve or stay level while both difference tasks decline by about 2 mIoU. Because the 27B model was trained for six epochs with a smaller warmup and no weight decay, and reached a higher validation loss than the 9B model (Table 8), this pattern should not be read as a property of scale alone. A matched schedule across sizes would be needed to separate the effect of capacity from the effect of optimization.

D.5 Why GRPO adds little on top of dCoT-SFT

GRPO learns from differences in reward among completions sampled for the same prompt. After dCoT-SFT, the policy already solves most prompts on most samples, so many groups are likely to contain several completions with similar IoU and therefore provide weak contrast. At the same time, the sampling diagnostic shows real headroom: at least one of eight samples succeeds on 84.4% of hard training prompts while the average sample succeeds on 67.7%. That this headroom does not translate into benchmark gains suggests that the bottleneck is not the absence of good samples but the signal used to prefer them. An IoU reward cannot distinguish a box that overlaps the target by chance from one that identifies the correct finding, which is consistent with the concept-grounding decline.

Rewards that also score semantic correctness, or that reward the image-selection step separately from localization, are natural next steps.

The development studies also carry a methodological lesson. An apparent 3 to 5 point regression after GRPO was caused by an exported image processor with a lower resolution limit, not by learning. A zero-learning-rate save-and-reload control exposed the problem immediately. We recommend such a control whenever a fine-tuned vision-language model is exported and re-evaluated, since preprocessing mismatches can easily be mistaken for training effects of the same size as the effects being studied.

D.6 What box conditioning implies

Both stages of the pipeline see the annotated box. This makes the rationales explanations of a known answer: they are aligned with the correct region by construction, but they do not record how a model would search for it. We chose this design so that every rationale is aligned with the correct region; a rationale written without the answer could justify an incorrect region and teach the student to reason fluently toward wrong answers. The cost is a risk of post-hoc rationalization. The inference-time controls address the part of this concern that matters for the trained model: its localization changes substantially when the rationale is removed or replaced, so the rationale is not ignored. They do not show that the rationale faithfully reflects the model's internal computation, and they cannot rule out that some of the text is decorative. Rationales generated without the answer and filtered by correctness, and multi-rater clinical review, would strengthen the corpus in future versions.

D.7 Practical takeaways

For practitioners building grounding models for multi-image clinical data, our results suggest four points. First, reasoning supervision is most valuable when the target is defined by a relation between images, such as change or correspondence across slices. Second, refinement should keep access to the images rather than relying on a caption. Third, gains should be checked per task, since reasoning supervision can hurt appearance-matching tasks while improving the average. Fourth, reward optimization after reasoning supervision should be evaluated with matched preprocessing, multiple seeds and rewards that credit correct identification, not only overlap.

E Limitations

First, both generation stages inherit model blind spots, and the annotated box conditions the final explanation; the rationales may therefore be post-hoc and should not be interpreted as faithful internal reasoning. The single-rater audit covers only 2,000 examples and lacks item-level labels, specialty metadata, error-category counts, and inter-rater agreement. Second, the 4B answer-only and dCoT-SFT runs differ in training stage, effective batch, optimizer implementation, and checkpoint selection. Their paired test analysis is a fixed-model comparison, not a causal or multi-seed estimate. Third, the SFT and test sets share image content in 147 test examples and a task-aware source series in 17; although no complete sequence or series-plus-box duplicates and a leakage-clean sensitivity analysis are reported, strict image- or study-level disjointness is not established. Fourth, evaluation is limited to MedSG-Bench and the CoT benefit is negative on several tasks. Finally, the GRPO findings apply only to the verified IoU+format reward and optimization configuration; they do not establish that GRPO is generally ineffective. The no-think and corrupted- reasoning controls are inference-time distribution shifts on one SFT checkpoint and do not by themselves establish rationale faithfulness or training-run causality. The tasks where reasoning supervision does not yet help behave very differently from one another, and a single IoU-based reward does not supervise them well. Future work should use matched multi-seed controls, patient/study-level splits, multi-specialist audits, and richer rewards.

F Broader Impacts

MedChain-R1 and MedChain-182K are intended for research use in medical AI. Positive impacts include improved interpretability of medical image grounding through explicit clinical reasoning chains Wei et al. (2022), which could support radiologist workflows in longitudinal scan review,

multi-view analysis, and cross-modal examination. Negative risks include potential over-reliance on model predictions in clinical settings without appropriate human oversight, and the possibility that errors in pseudo-CoT generation (Appendix A.2) propagate into model behavior. We strongly caution against deployment without rigorous clinical validation. MedChain-182K is constructed from publicly available benchmark data Yue et al. (2026) and does not contain identifiable patient information.

G Qualitative Case Studies

Figure 4 and Figures 5–9 show representative predictions of zero-shot Qwen3.5, MedSeq-Grounder and MedChain-R1. The examples were selected for illustration and are not a random sample.

H Extended Related Work

Medical vision-language models. Large-scale medical VLMs such as LLaVA-Med (Li et al., 2023), HuatuoGPT-Vision (Chen et al., 2024a) and MedGemma (Sellergren et al., 2025) perform well on medical VQA and report generation. FlemingVL (Shu et al., 2025) provides a unified framework for heterogeneous medical visual understanding across 2D, 3D and video modalities and serves as the Stage 1 teacher in our pipeline. General-purpose MLLMs including Qwen2.5-VL (Bai et al., 2025b), InternVL3 (Zhu et al., 2025) and the multi-image grounding model Migician (Li et al., 2025) serve as zero-shot baselines. None of these models is trained for sequential medical grounding, and none receives cross-image reasoning supervision.

Chain-of-thought distillation. Distilling reasoning from large models into smaller deployable ones (Guo et al., 2025; Wei et al., 2022) is a practical way to scale chain-of-thought supervision. In medical settings the central difficulty is aligning fluent clinical explanations with visual evidence across a sequence. In our pipeline FlemingVL-38B (Shu et al., 2025) supplies a preliminary visual rationale, and multimodal Qwen3.5-397B-A17B (Team, 2026) re-examines the original images before writing the final box-conditioned explanation. This differs from text-only distillation because the refiner keeps direct visual access.

Visual grounding architectures. Classical grounding models such as TransVG (Deng et al., 2021), MDETR (Kamath et al., 2021) and Grounding DINO (Liu et al., 2024) predict boxes from a single image and text query. Sequential grounding adds image selection and cross-image correspondence, which motivates generating an explicit comparison before the box.

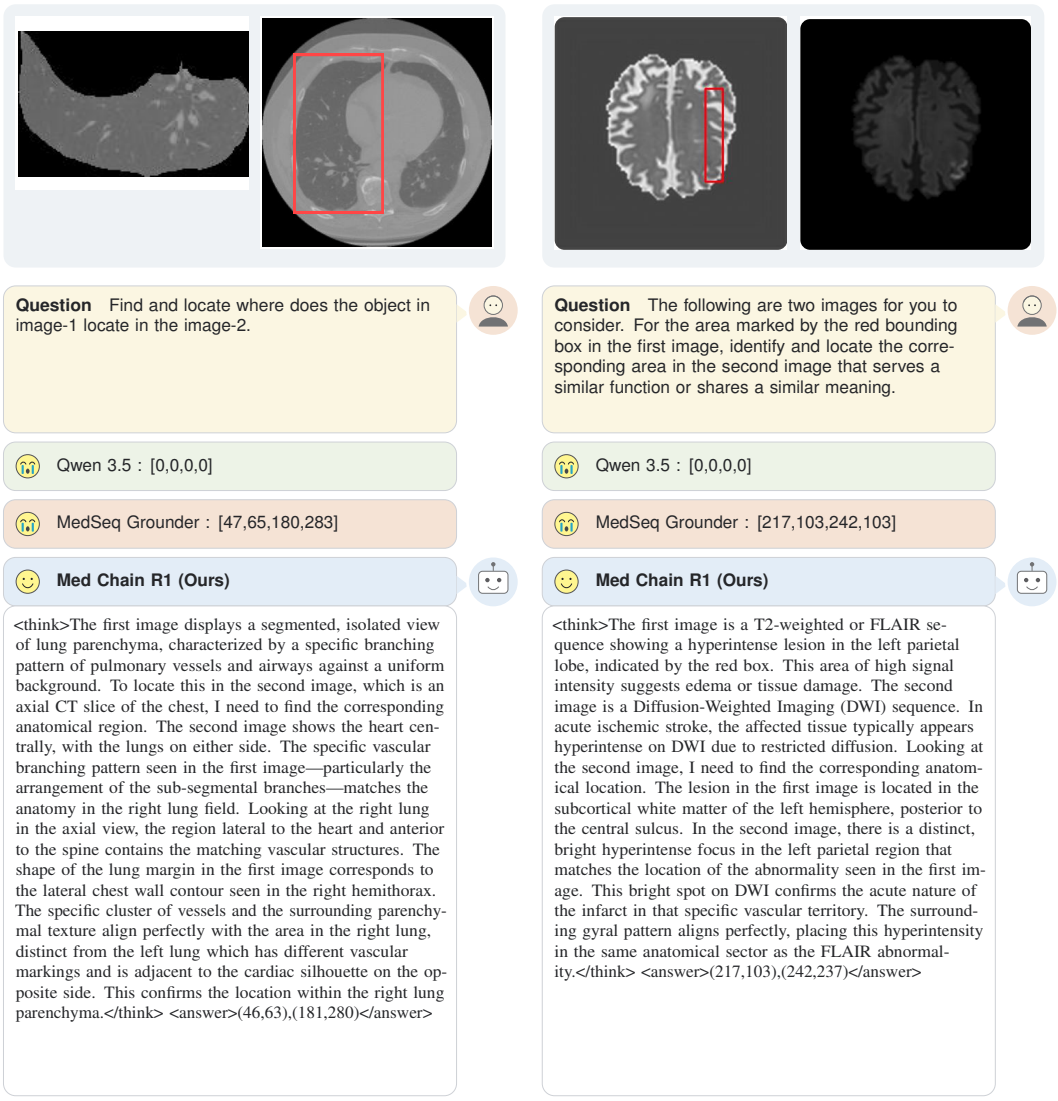
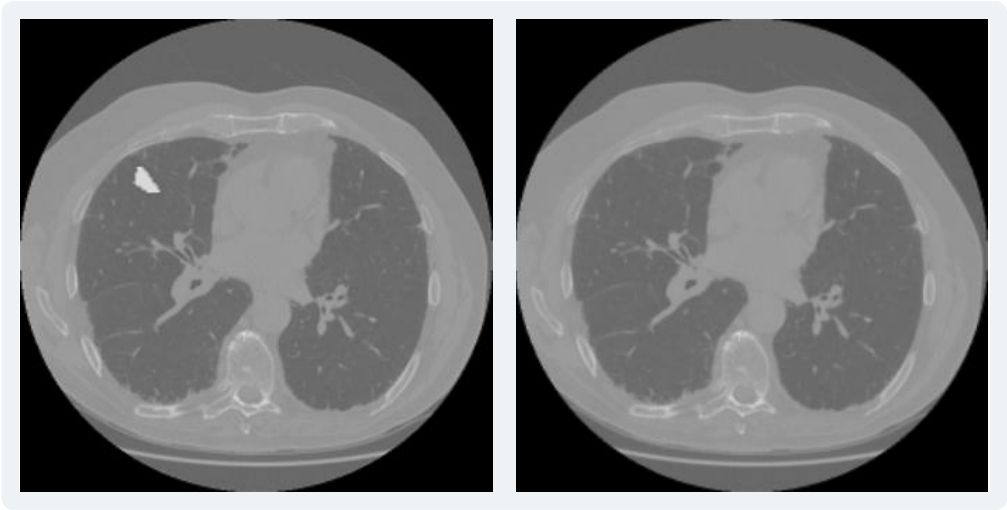


Figure 4: **Qualitative results of MEDCHAIN-R1 on sequential medical image grounding.** Two representative examples from MedSG-Bench: object tracking across CT slices (*left*) and cross-modal consistency grounding between FLAIR and DWI MRI sequences (*right*). For each example, zero-shot Qwen3.5 produces degenerate outputs ($[0, 0, 0, 0]$), MedSeq-Grounder predicts plausible but imprecise boxes, and MEDCHAIN-R1 generates structured clinical reasoning chains (`<think>`) that identify the correct anatomical correspondence before predicting precise bounding box coordinates (`<answer>`). The reasoning traces demonstrate cross-image analysis vascular pattern matching in CT and ischemic lesion localization via DWI hyperintensity in MRI that is absent in both baselines.



Question Compare these two images carefully and give me the coordinates of their difference.

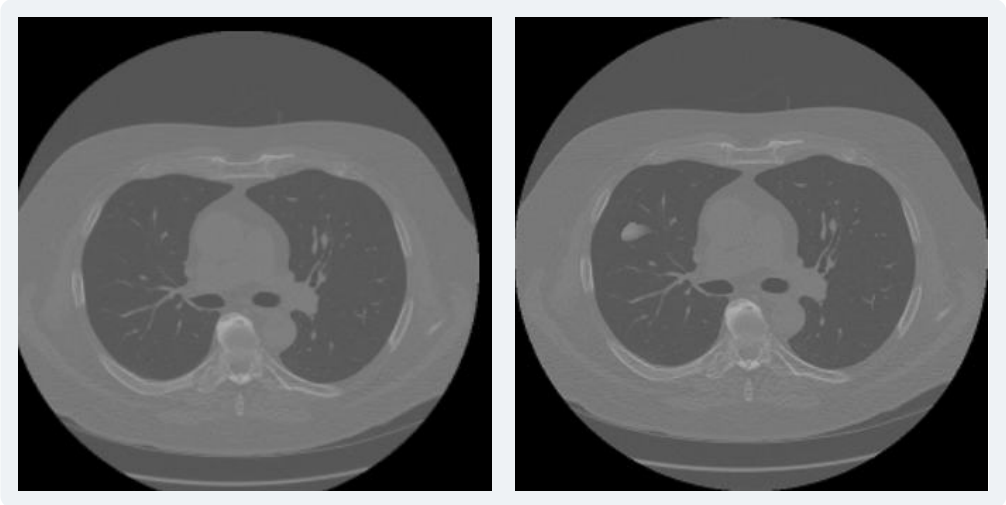
Qwen 3.5 : [0,0,0,0]

MedSeq Grounder : [79,105,98,120]

Med Chain R1 (Ours)

<think>The first image shows a distinct, high-density, irregular opacity in the right upper lobe, consistent with a pulmonary nodule or mass. In the second image, this specific opacity is absent, leaving only the normal vascular markings in that location. The region of interest must therefore isolate this focal lesion to capture the change. Is the change simply a shift in windowing? No, the surrounding mediastinal and bony structures remain identical in density and position, confirming a true anatomical difference rather than an artifact. The discrepancy is localized strictly to the area where the nodule was previously visualized.</think> <answer>(79,104),(99,123)</answer>

Figure 5: **Case Study 1: Registered Difference Grounding (CT).** Given two registered axial CT chest scans, the task is to localize the interval change between them. **Qwen3.5** completely fails, predicting `[0,0,0,0]` unable to detect any difference. **MedSeq-Grounder** predicts `[79,105,98,120]`, identifying the approximate region but with imprecise boundaries. **MEDCHAIN-R1 (Ours)** reasons that the first image shows a distinct high-density irregular opacity in the right upper lobe consistent with a pulmonary nodule, absent in the second image, rules out windowing artifact by confirming stable mediastinal and bony structures, and correctly localizes the nodule at `(79,104),(99,123)`. This case illustrates that correct difference grounding requires clinical concept identification distinguishing a true focal lesion from normal vascular markings rather than naive pixel comparison.



Question Compare these two images carefully and give me the coordinates of their real difference in the second image. Find it and locate it in the second image.

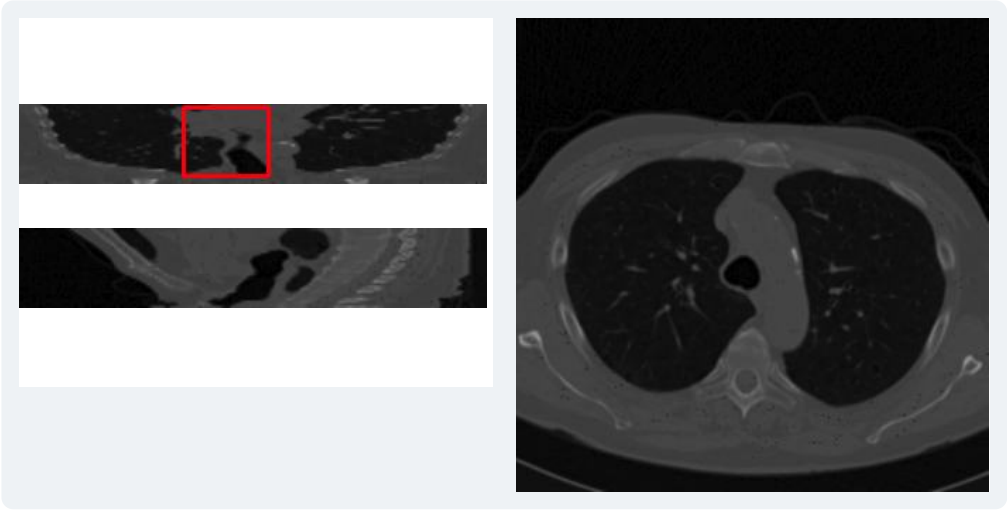
Qwen 3.5 : [0,0,0,0]

MedSeq Grounder : [75,145,96,153]

Med Chain R1 (Ours)

<think>The initial scan of the right lung field reveals a subtle, well-circumscribed opacity in the peripheral parenchyma that warrants closer inspection. Is this merely a vessel seen end-on or a true nodule? Comparing the two slices, the first image shows clear lung markings in this specific lateral segment, whereas the second image demonstrates a new, discrete, rounded density. This focal opacity stands out against the aerated lung background and lacks the tapering appearance of a normal bronchovascular bundle, confirming it as a pathological finding rather than an anatomical variant or artifact.</think> <answer>(75,145),(98,158)</answer>

Figure 6: **Case Study 2: Non-Registered Difference Grounding (CT).** Given two non-registered axial CT chest scans, the task is to localize the real difference in the second image. **Qwen3.5** again predicts `[0,0,0,0]`, failing completely to detect any change. **MedSeq-Grounder** predicts `[75,145,96,153]`, localizing the correct general region but with imprecise boundaries. **MEDCHAIN-R1 (Ours)** reasons that the right lung field shows a subtle well-circumscribed opacity in the peripheral parenchyma, questions whether it is a vessel seen end-on or a true nodule, confirms it as a new pathological finding by noting the absence of the tapering appearance of a normal bronchovascular bundle, and correctly localizes it at `(75,145),(98,158)`. Compared to Case 1 (registered CT), this harder non-registered setting requires the model to establish spatial correspondence without alignment cues, yet MEDCHAIN-R1's clinical reasoning correctly identifies and localizes the interval finding where both baselines fail.



Question These images share one object in common(the object marked with red bounding box in the first image(<|box_start|>(121,2),(181,50)<|box_end|>). Recognize and locate this object in the second image.

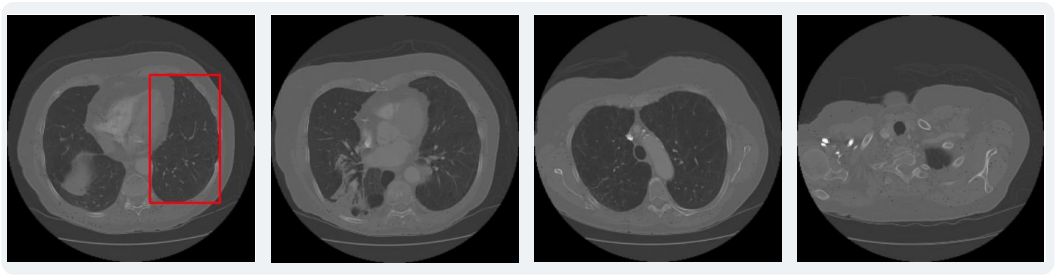
Qwen 3.5 : [121,2,181,50]

MedSeq Grounder : [139,16,194,54]

Med Chain R1 (Ours)

<think>The first image shows a coronal slice where the red box highlights a segment of the central airway, specifically the trachea or proximal main bronchus, appearing as a dark, air-filled lumen surrounded by soft tissue. To find the corresponding structure in the second image, which is a sagittal view, I need to look for the vertical course of the trachea. In the sagittal plane, the trachea presents as a distinct, air-density column anterior to the spine and esophagus. The region identified contains this vertical air-filled structure, consistent with the tracheal lumen seen in the coronal view. It is positioned anterior to the vertebral column and maintains the expected tubular morphology of the central airway, distinguishing it from the esophagus or vascular structures nearby.</think> <answer>(140,16),(195,55)</answer>

Figure 7: **Case Study 3: Crossmodal Grounding (CT).** Given a coronal CT slice with a red-boxed reference region and an axial CT slice, the task is to localize the same anatomical structure in the second image across views. **Qwen3.5** simply copies the reference box coordinates `[121,2,181,50]`, failing to perform any cross-view reasoning. **MedSeq-Grounder** predicts `[139,16,194,54]`, identifying the approximate region but without view-consistent reasoning. **MEDCHAIN-R1 (Ours)** identifies the red-boxed region in the coronal view as the trachea or proximal main bronchus a dark air-filled lumen surrounded by soft tissue reasons that in the sagittal axial view the trachea presents as a vertical air-density column anterior to the spine and esophagus, distinguishes it from nearby vascular structures by its tubular morphology, and correctly localizes it at `(140,16),(195,55)`. This case demonstrates that crossmodal grounding requires explicit anatomical concept transfer across imaging planes, which MEDCHAIN-R1's clinical reasoning chain uniquely provides.



Question These images share one object in common(the object marked with red bounding box in the first image(<|box_start|>(193,81),(288,255)<|box_end|>). Recognize and locate this object in the second image.

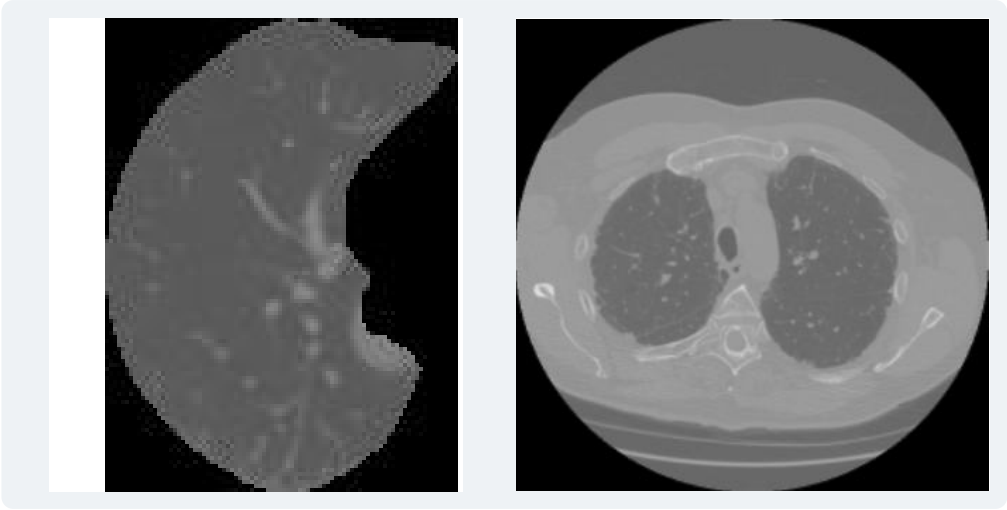
Qwen 3.5 : [0,0,0,0]

MedSeq Grounder : [187,83,284,206]

Med Chain R1 (Ours)

<think>The first image shows a bounding box over the left lung field, specifically highlighting a region of lung parenchyma. To track this object in the second image, I need to identify the corresponding anatomical slice. The first image is at a lower thoracic level, showing the heart and lower lobes. The second image is slightly superior, showing the main pulmonary artery and bronchi more clearly. The target is the left lung. In the second image, the left lung is the large, dark, air-filled structure on the right side of the image (patient's left). The region corresponding to the previous slice's lung parenchyma is the main body of the left lung field. I should look for the continuity of the lung tissue itself. The box in the second image should encompass the equivalent area of the left lung parenchyma. The region on the patient's left (right side of the image) is the correct anatomical location. The specific area matches the lung field seen in the first image, maintaining the lateral position relative to the mediastinum.</ think> <answer>(187,83),(284,265)</answer>

Figure 8: **Case Study 4: Object Tracking (CT).** Given four sequential axial CT slices with a red-boxed reference region in the first image, the task is to localize the same object in the second image across slice levels. **Qwen3.5** predicts `[0,0,0,0]`, failing entirely. **MedSeq-Grounder** predicts `[187,83,284,206]`, identifying the correct region but with imprecise inferior boundary. **MEDCHAIN-R1 (Ours)** reasons that the red box in the first image highlights left lung parenchyma at a lower thoracic level showing the heart and lower lobes, identifies that the second image is a slightly superior slice showing the main pulmonary artery and bronchi, tracks the left lung as the large air-filled structure on the patient's left, and correctly localizes the corresponding parenchymal region at `(187,83),(284,265)` by maintaining lateral position relative to the mediastinum across slice levels. This case highlights that object tracking across CT slices requires anatomical continuity reasoning understanding how structures evolve across imaging planes which neither baseline can perform.



Question Find and locate where does the object in image-1 locate in the image-2.

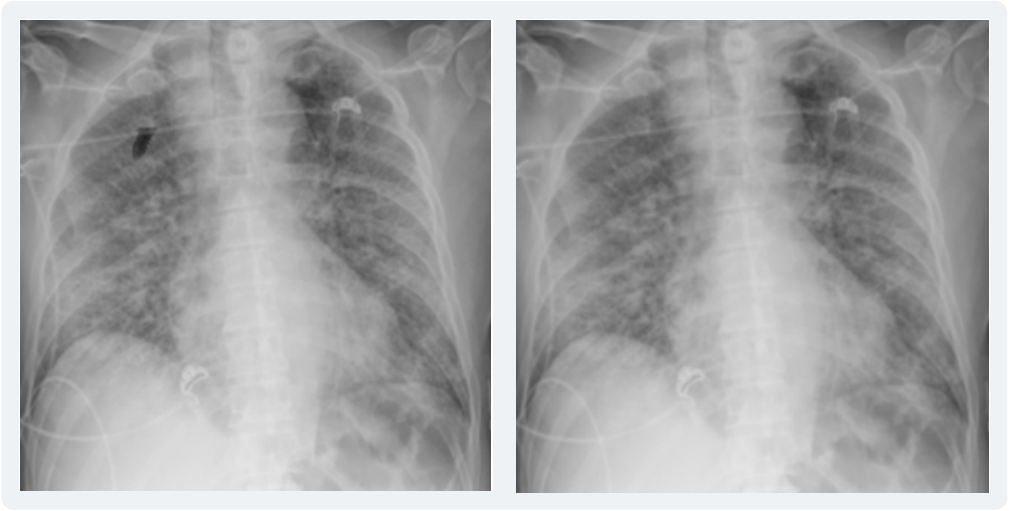
Qwen 3.5 : [0,0,0,0]

MedSeq Grounder : [44,99,158,250]

Med Chain R1 (Ours)

<think>The first image displays a segmented, isolated view of a pulmonary nodule, characterized by a distinct, rounded opacity with relatively well-defined margins against a dark background. To locate this in the second image, which is an axial CT slice of the chest, I need to scan the lung parenchyma for a matching focal lesion. Looking at the right lung field (left side of the image), there is a solitary, hyperdense nodule situated in the posterior aspect of the upper lobe. This lesion stands out against the surrounding aerated lung tissue and matches the morphology seen in the reference crop. Is this the only nodule? There are vascular markings nearby, but this specific opacity is distinct from the branching pulmonary vessels, confirming it as the target pathology.</ think> <answer>(44,99),(158,254)</answer>

Figure 9: **Case Study 5: Concept Localization (CT).** Given a segmented isolated crop of a pulmonary nodule (image 1) and a full axial CT chest slice (image 2), the task is to localize the same nodule in the full scan. **Qwen3.5** predicts `[0,0,0,0]`, failing completely. **MedSeq-Grounder** predicts `[44,99,158,250]`, spanning an overly large region that includes surrounding lung tissue. **MEDCHAIN-R1 (Ours)** identifies the reference crop as a rounded well-defined opacity consistent with a pulmonary nodule, scans the right lung field for a matching focal lesion, locates a solitary hyperdense nodule in the posterior aspect of the upper lobe, distinguishes it from nearby branching pulmonary vessels by its distinct morphology, and correctly tightens the localization to `(44,99),(158,254)`. This case is the hardest task type for simple RL concept localization requires matching a pathological finding across heterogeneous imaging contexts, explaining the severe semantic drift (-6.70 Acc@0.5) observed under IoU-only GRPO.



Question Compare these two images carefully and give me the coordinates of their difference.

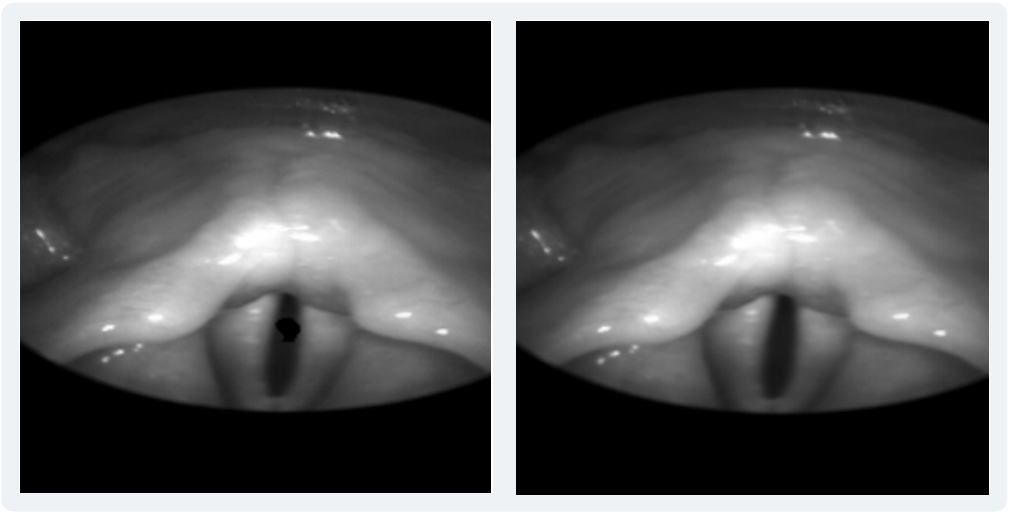
Qwen 3.5 : [0,0,0,0]

MedSeq Grounder : [80,77,95,94]

Med Chain R1 (Ours)

<think>The upper right lung field shows a distinct, well-circumscribed radiolucent area that stands out against the surrounding pulmonary parenchyma. This focal lucency suggests a localized abnormality, such as a bulla or a cavity, which is not mirrored in the contralateral lung or the adjacent lung zones. Is this simply a projectional artifact or a true pathological finding? The sharp definition and density contrast confirm it as a significant structural change rather than overlapping soft tissue or rib shadow.</think> <answer>(79,77),(97,99)</answer>

Figure 10: **Case Study 6: Registered Difference Grounding (Chest X-ray).** Given two registered chest radiographs, the task is to localize the difference between them. **Qwen3.5** predicts `[0,0,0,0]`, an empty box. **MedSeq-Grounder** predicts `[80,77,95,94]`. **MEDCHAIN-R1 (Ours)** identifies a well-circumscribed radiolucent area in the upper right lung field, checks whether it could be a projectional artifact or overlapping rib shadow, and predicts `(79,77),(97,99)`.



Question Compare these two images carefully and give me the coordinates of their difference.

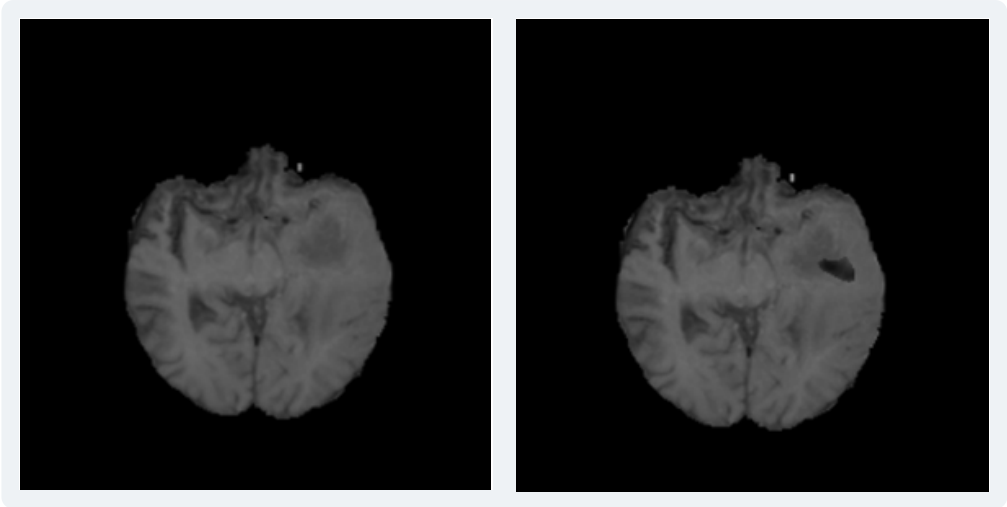
Qwen 3.5 : [450,550,550,750]

MedSeq Grounder : [181,211,198,226]

Med Chain R1 (Ours)

<think>The first image shows a distinct, dark, rounded opacity located centrally within the glottic airway, just inferior to the anterior commissure. This structure is absent in the second image, where the corresponding area appears as a clear, dark vertical slit representing the open vocal folds and airway lumen. The discrepancy is localized to this specific focal lesion; the surrounding vocal fold mucosa and the anterior commissure itself appear structurally consistent between the two frames, suggesting the change is isolated to this intraluminal object rather than a generalized movement or lighting shift. Is this dark spot an artifact or a true pathology? Given its well-defined borders and central location, it likely represents a foreign body or a specific lesion like a blood clot or tumor that has been removed or resolved in the second frame. The key visual cue is the presence of this discrete, high-contrast density in the first image that completely disappears in the second, leaving the native airway geometry exposed. Therefore, the region of interest must tightly encompass this specific opacity to capture the change accurately.</think> <answer>(180,211),(200,227)</answer>

Figure 11: **Case Study 7: Registered Difference Grounding (Laryngoscopy).** Given two registered laryngoscopic frames, the task is to localize the difference between them. **Qwen3.5** predicts `[450,550,550,750]`. **MedSeq-Grounder** predicts `[181,211,198,226]`. **MEDCHAIN-R1 (Ours)** describes a dark, rounded intraluminal opacity in the glottic airway that is present in the first frame and absent in the second, notes that the surrounding vocal fold mucosa is unchanged, and predicts `(180,211),(200,227)`.



Question Compare these two images carefully and give me the coordinates of their real difference in the second image. Find it and locate it in the second image.

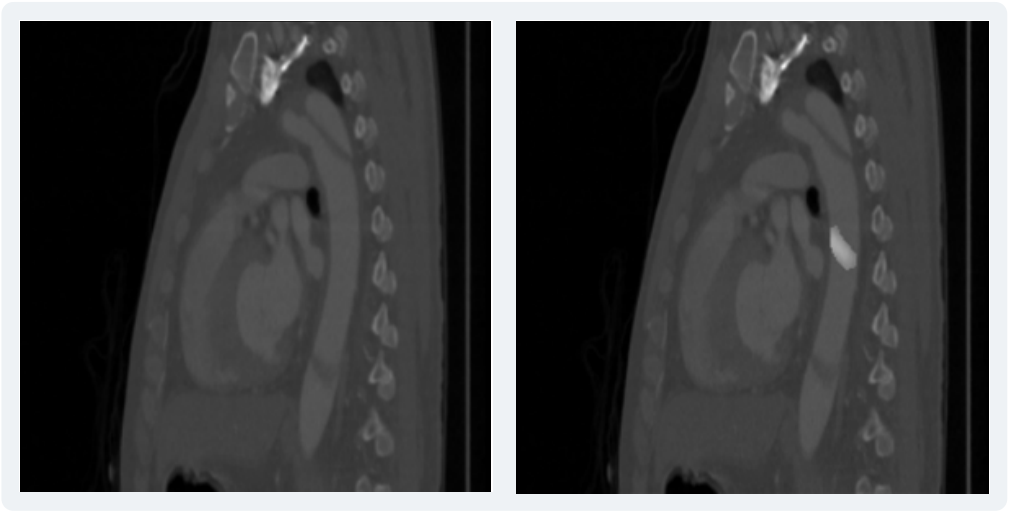
Qwen 3.5 : [0,0,0,0]

MedSeq Grounder : [214,169,240,173]

Med Chain R1 (Ours)

<think>The first image shows a relatively homogeneous signal intensity in the left temporal lobe region, consistent with normal parenchyma or perhaps subtle edema, but no distinct focal lesion is clearly demarcated. In the second image, a new, well-defined hypointense area appears in the left temporal lobe. This dark, irregular focus suggests a cystic component, necrosis, or a fluid-filled cavity that was not present or visible in the initial scan. Is this just a shift artifact? No, the surrounding sulci and gyri remain relatively stable in position, indicating this is a true local change rather than global misalignment. The specific area of interest is this new hypointensity, which stands out against the surrounding gray matter. It represents a distinct pathological development, likely tumor progression or a new cystic change, making it the primary finding to localize.</think> <answer>(213,169),(240,187)</answer>

Figure 12: **Case Study 8: Non-Registered Difference Grounding (Brain MRI).** Given two non-registered brain MRI slices, the task is to localize the real difference in the second image. **Qwen3.5** predicts `[0,0,0,0]`, an empty box. **MedSeq-Grounder** predicts `[214,169,240,173]`. **MEDCHAIN-R1 (Ours)** identifies a new hypointense focus in the left temporal lobe, rules out a global shift by noting that the surrounding sulci and gyri are stable, and predicts `(213,169),(240,187)`.



Question Compare these two images carefully and give me the coordinates of their real difference in the second image. Find it and locate it in the second image.

Qwen 3.5 : [600,530,660,600]

MedSeq Grounder : [223,145,241,160]

Med Chain R1 (Ours)

<think>The sagittal CTA view shows the thoracic aorta and adjacent mediastinal structures. In the second image, there is a distinct, high-density, linear structure located within the lumen of the descending thoracic aorta that is absent in the first image. This hyperdense object appears to be an endovascular stent graft or a similar intraluminal device, characterized by its bright attenuation relative to the surrounding contrast-filled blood pool. Is this simply calcification? No, the morphology is too regular and linear, suggesting an artificial implant rather than irregular plaque. The location corresponds to the expected path of the descending aorta, just posterior to the heart and anterior to the vertebral column. The surrounding lung parenchyma and vertebral bodies appear stable, making this new intraluminal density the primary interval change. The specific area of interest is where this bright, foreign material is clearly visualized within the vessel lumen.</think>
<answer>(222,146),(241,177)</answer>

Figure 13: **Case Study 9: Non-Registered Difference Grounding (CT Angiography).** Given two non-registered sagittal CT angiography images, the task is to localize the real difference in the second image. **Qwen3.5** predicts `[600,530,660,600]`. **MedSeq-Grounder** predicts `[223,145,241,160]`. **MEDCHAIN-R1 (Ours)** identifies a new linear hyperdense structure within the lumen of the descending thoracic aorta, distinguishes it from calcification by its regular morphology, and predicts `(222,146),(241,177)`.





Question These images share one object in common (the object marked with red bounding box in the first image <|box_start|>(37,151),(63,180)<|box_end|>). Recognize and locate this object in the second image.

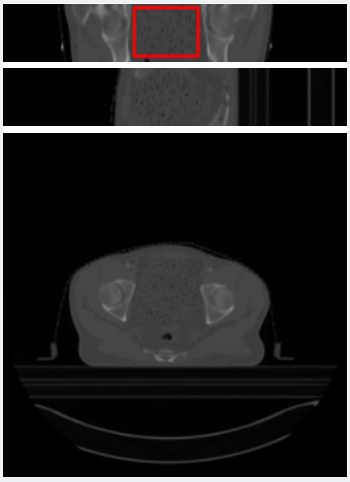
Qwen 3.5 : [37,151,63,180]

MedSeq Grounder : [49,190,74,212]

Med Chain R1 (Ours)

<think>The CC view shows a focal, irregular density in the posterior breast tissue, marked by the red box. This lesion appears somewhat spiculated or ill-defined, distinguishing it from the surrounding fibroglandular tissue. When correlating with the MLO view, I need to find a corresponding mass in the upper outer quadrant or posterior aspect, as the CC view places it laterally. Looking at the MLO image, there is a distinct, dense mass located in the upper posterior region, just anterior to the pectoral muscle shadow. Its density and irregular margins match the lesion seen on the CC view. The position is consistent with the posterior location noted in the first image. Is there any other structure that could be confused? The nipple is visible inferiorly, but its shape and location are distinct. The dense area in the upper posterior breast is the only structure with matching morphology and relative depth.</think> <answer>(62,153),(95,185)</answer>

Figure 14: **Case Study 10: Multi-view Grounding (Mammography).** Given a craniocaudal (CC) view with a red-boxed lesion and a mediolateral oblique (MLO) view, the task is to localize the same lesion in the MLO view. **Qwen3.5** predicts `[37,151,63,180]`, copying the reference box. **MedSeq-Grounder** predicts `[49,190,74,212]`. **MEDCHAIN-R1 (Ours)** relates the posterior position of the lesion on the CC view to the MLO projection, locates a dense, irregular mass anterior to the pectoral muscle shadow, and predicts `(62,153),(95,185)`.



Question These images share one object in common(the object marked with red bounding box in the first image(<|box_start|>(128,4),(190,50)<|box_end|>). Recognize and locate this object in the third image.

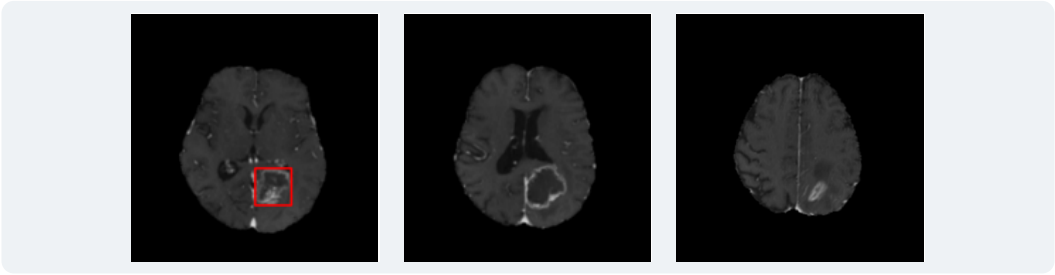
Qwen 3.5 : [128,4,190,50]

MedSeq Grounder : [131,126,187,148]

Med Chain R1 (Ours)

<think>The first image shows a coronal view of the pelvis, where the red box highlights the bladder, characterized by its central location and fluid density, situated between the iliac bones. To find the corresponding structure in the third image, which is an axial slice, I need to look for the same fluid-filled organ in the midline of the pelvis. In the axial view, the bladder appears as a hypodense, roughly rounded structure anterior to the rectum and between the pelvic bones. The region bounded by the box in the third image captures this central, low-attenuation area perfectly, matching the anatomical position and appearance of the bladder seen in the coronal reference. Is there any other structure that could be confused here? The surrounding bowel loops have different densities and textures, confirming this central fluid collection is indeed the bladder.</think> <answer>(128,125),(190,186)</answer>

Figure 15: **Case Study 11: Multi-view Grounding (CT).** Given a coronal pelvic CT image with the bladder marked by a red box, the task is to localize the same structure in the third image, an axial slice. **Qwen3.5** predicts `[128,4,190,50]`, copying the reference box. **MedSeq-Grounder** predicts `[131,126,187,148]`. **MEDCHAIN-R1 (Ours)** identifies the boxed structure as the bladder, searches the axial slice for the midline fluid-density organ anterior to the rectum, and predicts `(128,125),(190,186)`.



Question These images share one object in common(the object marked with red bounding box in the first image(<|box_start|>(167,209),(216,259)<|box_end|>). Recognize and locate this object in the third image.

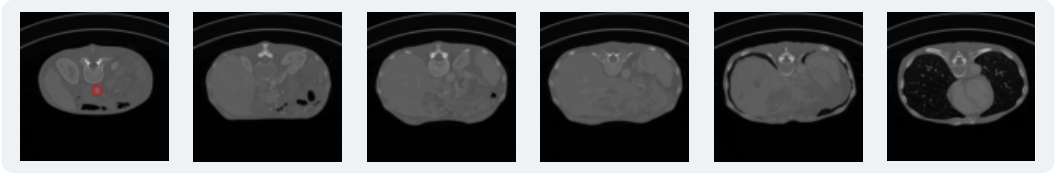
Qwen 3.5 : [167,209,216,259]

MedSeq Grounder : [178,197,214,232]

Med Chain R1 (Ours)

<think>The first image shows a heterogeneously enhancing lesion in the left parieto-occipital region, characterized by a central area of necrosis or cystic change surrounded by a thick, irregular rim of enhancement. As we track this pathology superiorly to the third image, the lesion maintains its location in the left posterior hemisphere. The specific target in the third image corresponds to the superior aspect of this same mass. It presents as a distinct, hyperintense nodular focus against the darker background of the brain parenchyma. While there is some surrounding edema, the boxed region specifically isolates the solid, enhancing tumor component which remains the most conspicuous feature, distinguishing it from the adjacent vasogenic edema or normal cortical structures.</think> <answer>(178,226),(204,254)</answer>

Figure 16: **Case Study 12: Object Tracking (Contrast-enhanced MRI).** Given three brain MRI slices with an enhancing lesion marked by a red box in the first image, the task is to localize the same lesion in the third image. **Qwen3.5** predicts `[167,209,216,259]`, copying the reference box. **MedSeq-Grounder** predicts `[178,197,214,232]`. **MEDCHAIN-R1 (Ours)** follows the left posterior lesion superiorly, separates the solid enhancing component from the surrounding vasogenic edema, and predicts `(178,226),(204,254)`.



Question These images share one object in common(the object marked with red bounding box in the first image(<|box_start|>(166,168),(182,185)<|box_end|>). Recognize and locate this object in the fourth image.

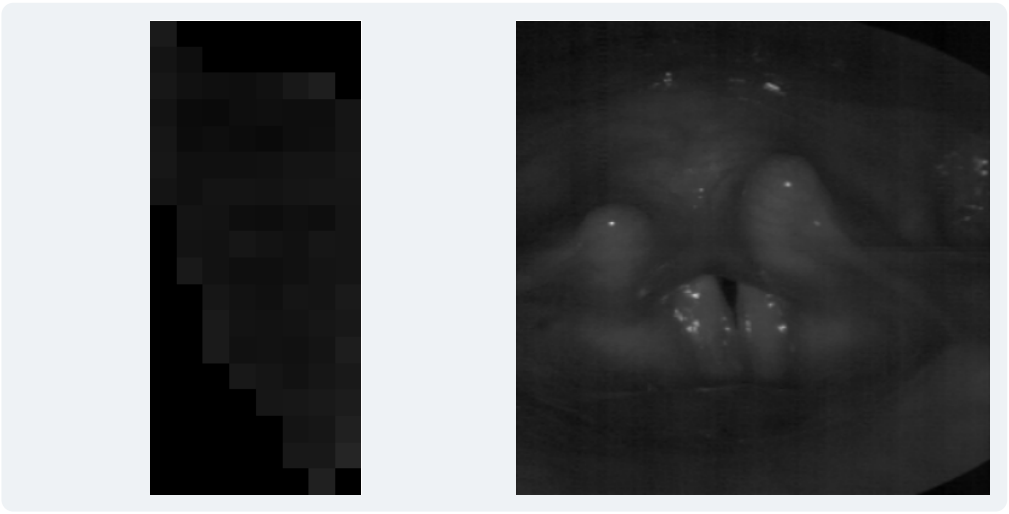
Qwen 3.5 : [166,168,182,185]

MedSeq Grounder : [164,150,197,173]

Med Chain R1 (Ours)

<think>The target structure is the abdominal aorta, visualized as a circular, soft-tissue density vessel located anterior to the vertebral body and slightly to the left of the midline. In the reference slice, it sits just anterior to the spine, distinct from the inferior vena cava which would be to its right. As we track superiorly to the fourth image, the aorta maintains its characteristic round morphology and pre-vertebral position. It is crucial to distinguish this from the adjacent esophagus or other mediastinal structures; however, in this abdominal slice, the vessel remains consistent in size and location relative to the vertebral column. The boxed region correctly captures this vascular structure, ensuring continuity with the initial finding.</think> <answer>(182,126),(203,148)</answer>

Figure 17: **Case Study 13: Object Tracking (Abdominal CT).** Given a sequence of axial CT slices with the abdominal aorta marked by a red box in the first image, the task is to localize the same structure in the fourth image. **Qwen3.5** predicts `[166,168,182,185]`, copying the reference box. **MedSeq-Grounder** predicts `[164,150,197,173]`. **MEDCHAIN-R1 (Ours)** tracks the round prevertebral vessel across slices, distinguishes it from the inferior vena cava and the esophagus, and predicts `(182,126),(203,148)`.



Question Find and locate where does the object in image-1 locate in the image-2.

Qwen 3.5 : [0,0,250,400]

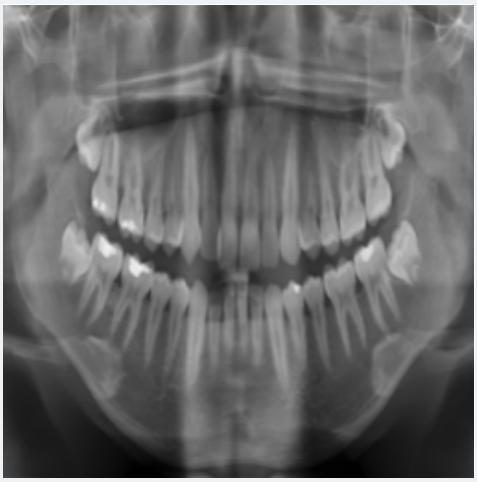
MedSeq Grounder : [146,183,155,207]

Med Chain R1 (Ours)

<think>The first image is a low-resolution, cropped view showing a dark, somewhat linear structure against a lighter background, which is characteristic of the glottic opening or a vocal fold edge seen in endoscopy. Looking at the second image, which is a full endoscopic view of the larynx, I need to find the corresponding anatomical feature. The larynx shows the epiglottis superiorly and the arytenoids posteriorly, with the vocal folds meeting in the midline. The specific shape in the first image—a dark slit or gap—strongly suggests the glottis itself, the space between the vocal cords. Scanning the central region of the larynx in the second image, there is a distinct, dark, vertical slit where the true vocal folds approximate. This area corresponds to the rima glottidis. The surrounding tissue is the brighter, reflective mucosa of the vocal folds. The cropped image matches the texture and contrast of this central glottic opening. While the arytenoids are nearby posteriorly, they are more bulbous and rounded, unlike the linear, dark appearance of the glottic chink. Therefore, the region of interest must be the glottic opening itself, located centrally between the vocal folds.</think> <answer>(146,183),(156,204)</answer>

Figure 18: **Case Study 14: Concept Grounding (Laryngoscopy).** Given a cropped endoscopic region (image 1) and a full laryngoscopic view (image 2), the task is to localize the cropped content in the full view. **Qwen3.5** predicts `[0,0,250,400]`. **MedSeq-Grounder** predicts `[146,183,155,207]`. **MEDCHAIN-R1 (Ours)** matches the dark linear structure in the crop to the glottic opening between the vocal folds, distinguishes it from the more rounded arytenoids, and predicts `(146,183),(156,204)`.





Question Find and locate where does the object in image-1 locate in the image-2.

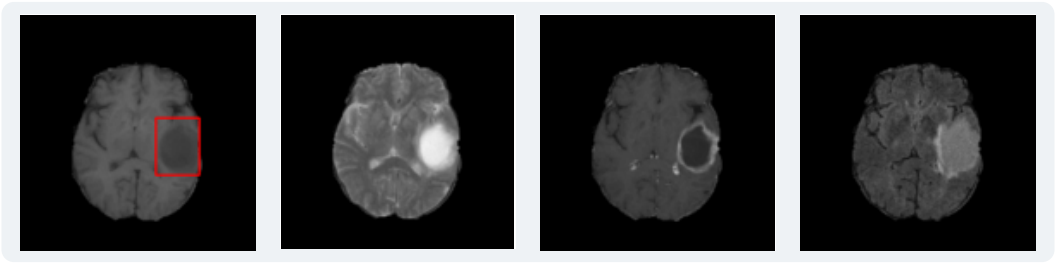
Qwen 3.5 : [100,200,900,850]

MedSeq Grounder : [40,65,307,238]

Med Chain R1 (Ours)

<think>The first image displays a segmented, high-contrast view of the maxillary and mandibular dentition, essentially a "mask" of the teeth. To locate this in the second image, which is a standard panoramic radiograph, I need to find the corresponding full set of teeth within the dental arches. The region of interest should encompass the entire visible dentition, including the crowns and roots, while excluding the surrounding maxillary sinus or mandibular ramus structures that are not part of the dental segment itself. Looking at the panoramic view, the central area containing the full arch of upper and lower teeth is the most prominent feature matching the segmentation. The boxed area captures the complete dental arcade, from the left posterior region across to the right posterior region, including both the maxillary and mandibular sets. This matches the content of the first image perfectly, as it isolates the teeth from the rest of the skeletal anatomy. Other regions would either miss parts of the dentition or include excessive non-dental structures like the temporomandibular joints or nasal cavity, which are not present in the reference mask.</think> <answer>(40,65),(299,274)</answer>

Figure 19: **Case Study 15: Concept Grounding (Dental Panoramic Radiograph).** Given a segmentation-style mask of the dentition (image 1) and a panoramic radiograph (image 2), the task is to localize the masked content in the radiograph. **Qwen3.5** predicts `[100,200,900,850]`. **MedSeq-Grounder** predicts `[40,65,307,238]`. **MEDCHAIN-R1 (Ours)** matches the mask to the full maxillary and mandibular dental arcade, excludes the sinus, mandibular ramus and temporomandibular joints, and predicts `(40,65),(299,274)`.



Question The following are four images for you to consider. For the area marked by the red bounding box in the first image, identify and locate the corresponding area in the third image that serves a similar function or shares a similar meaning.

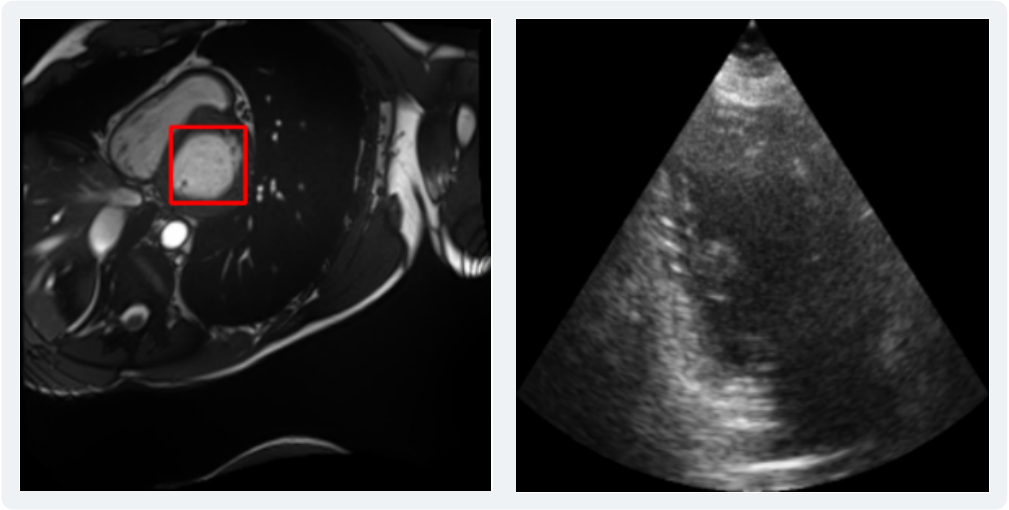
Qwen 3.5 : [650,400,780,650]

MedSeq Grounder : [198,147,255,147]

Med Chain R1 (Ours)

<think>The first image is a T1-weighted scan showing a hypointense mass in the left hemisphere, likely involving the temporal or parietal lobe. The third image is a T1-weighted post-contrast (T1ce) scan. In T1ce, enhancing tumors typically show a bright rim due to blood-brain barrier breakdown. Looking at the third image, there is a distinct ring-enhancing lesion in the same anatomical location as the hypointensity in the first image. The center of this ring is necrotic and non-enhancing, appearing dark, while the rim is bright. The bounding box encompasses this entire pathological structure—the necrotic core and the enhancing rim—which corresponds to the tumor mass seen in the first image. Is this the only lesion? Yes, it's the dominant pathology. Does the location match? Yes, it's in the left hemisphere, lateral to the midline structures. The box captures the full extent of this ring-enhancing tumor, which is the cross-modal equivalent of the mass effect seen in the first image.</think> <answer>(193,147),(255,228)</answer>

Figure 20: **Case Study 16: Cross-modal Grounding (Brain MRI, T1 to T1ce).** Given four MRI sequences with a hypointense mass marked by a red box on the T1-weighted image, the task is to localize the corresponding region in the third image, a post-contrast T1 (T1ce) scan. **Qwen3.5** predicts `[650,400,780,650]`. **MedSeq-Grounder** predicts `[198,147,255,147]`, a zero-height box. **MEDCHAIN-R1 (Ours)** relates the T1 hypointensity to a ring-enhancing lesion with a necrotic core at the same location on T1ce, and predicts `(193,147),(255,228)`.



Question The following are two images for you to consider. For the area marked by the red bounding box in the first image, identify and locate the corresponding area in the second image that serves a similar function or shares a similar meaning.

Qwen 3.5 : [0,0,0,0]

MedSeq Grounder : [97,60,268,154]

Med Chain R1 (Ours)

<think>The first image is an axial MRI showing the heart, where the red box highlights the left ventricular cavity, identifiable by its central location and fluid signal intensity relative to the surrounding myocardium. The second image is an apical 2-chamber echocardiogram. I need to find the corresponding anatomical structure, which is the left ventricle. In the ultrasound, the left ventricle appears as the large, hypoechoic chamber occupying the majority of the sector, tapering towards the apex. The region bounded by the endocardial border represents the blood pool, which corresponds directly to the cavity seen in the MRI. The surrounding brighter echoes represent the myocardial walls and pericardium. Therefore, the box should encompass this main ventricular chamber.</think> <answer>(119,88),(223,249)</answer>

Figure 21: **Case Study 17: Cross-modal Grounding (Cardiac MRI to Echocardiography).** Given an axial cardiac MRI image with the left ventricular cavity marked by a red box and an apical two-chamber echocardiogram, the task is to localize the corresponding region in the echocardiogram. **Qwen3.5** predicts `[0,0,0,0]`, an empty box. **MedSeq-Grounder** predicts `[97,60,268,154]`. **MEDCHAIN-R1 (Ours)** identifies the boxed region as the left ventricular cavity, locates the corresponding hypoechoic chamber bounded by the endocardial border in the ultrasound, and predicts `(119,88),(223,249)`.



Question You are given a source image followed by its several regions. Please locate the third region picture in the source image.

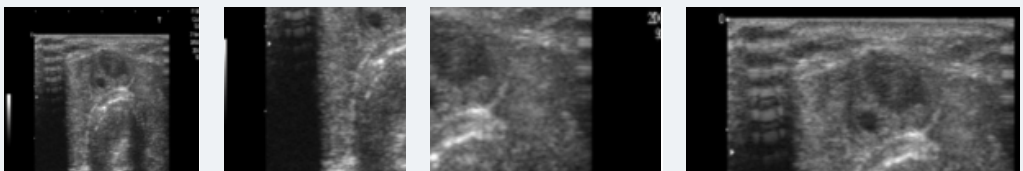
Qwen 3.5 : [0,0,512,512]

MedSeq Grounder : [0,113,308,264]

Med Chain R1 (Ours)

<think>The target region captures the central pigmented lesion, characterized by a reticular pattern of brownish pigment network. I need to distinguish this from the surrounding skin and surgical markings. The area shows a distinct transition from the darker, dense pigmentation in the upper left to a lighter, hypopigmented area in the lower right, which matches the specific gradient seen in the third crop. The presence of blue surgical ink markings at the periphery helps orient the location relative to the lesion's borders.</think> <answer>(16,91),(322,255)</answer>

Figure 22: **Case Study 18: Patch Grounding (Dermoscopy).** Given a dermoscopic source image followed by several of its regions, the task is to localize the third region in the source image. **Qwen3.5** predicts `[0,0,512,512]`, the whole image. **MedSeq-Grounder** predicts `[0,113,308,264]`. **MEDCHAIN-R1 (Ours)** uses the gradient from dense pigment network to a lighter hypopigmented area, together with the peripheral surgical ink markings, to place the region, and predicts `(16,91),(322,255)`.



Question You are given a source image followed by its several regions. Please locate the second region picture in the source image.

Qwen 3.5 : [0,0,0,0]

MedSeq Grounder : [96,85,304,177]

Med Chain R1 (Ours)

<think>The target region contains a distinct, hypoechoic nodule with somewhat irregular margins located within the thyroid parenchyma. Just inferior to this nodule, there is a prominent, curvilinear hyperechoic structure, likely representing the tracheal cartilage or a calcified interface, which serves as a key anatomical landmark. The texture within the box shows a heterogeneous echotexture typical of thyroid tissue, contrasting with the anechoic strap muscles anteriorly. Is this the only hypoechoic area? No, but the specific relationship between the nodule and the underlying bright, arc-like reflector is unique to this section of the gland. The surrounding tissue fades into the more homogeneous muscle layers laterally, confirming this central location within the thyroid lobe. The specific configuration of the nodule sitting just above the bright interface is the definitive cue.</think> <answer>(162,66),(325,210)</answer>

Figure 23: **Case Study 19: Patch Grounding (Thyroid Ultrasound).** Given a thyroid ultrasound source image followed by several of its regions, the task is to localize the second region in the source image. **Qwen3.5** predicts `[0,0,0,0]`, an empty box. **MedSeq-Grounder** predicts `[96,85,304,177]`. **MEDCHAIN-R1 (Ours)** anchors the region on a hypoechoic nodule sitting above a curvilinear hyperechoic interface, and predicts `(162,66),(325,210)`.

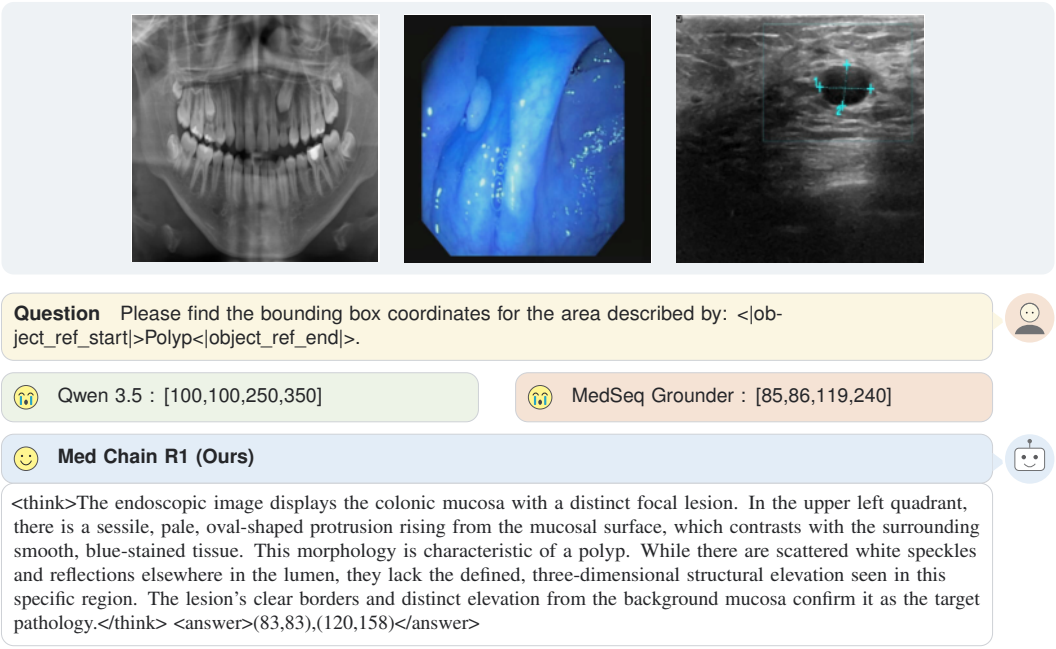
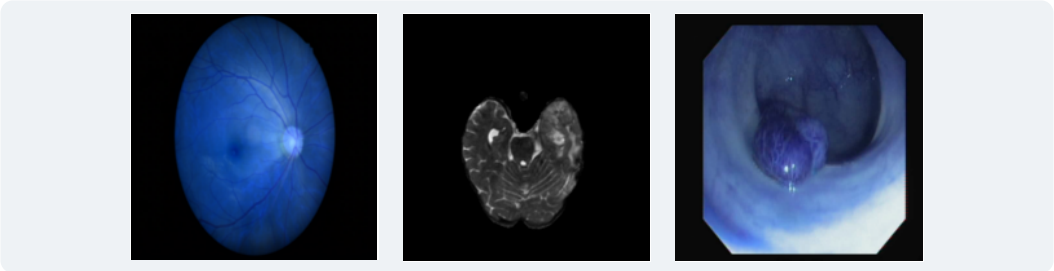


Figure 24: **Case Study 20: Referring Grounding (Colonoscopy).** Given the referring expression *Polyp*, the task is to localize the described region. **Qwen3.5** predicts `[100,100,250,350]`. **MedSeq-Grounder** predicts `[85,86,119,240]`. **MEDCHAIN-R1 (Ours)** describes a sessile, pale, oval protrusion in the upper left quadrant of the colonoscopic image, distinguishes it from scattered specular reflections, and predicts `(83,83),(120,158)`.



Question Please find the bounding box coordinates for the area described by: <|object_ref_start|>OpticCup<|object_ref_end|>.

Qwen 3.5 : [600,350,800,550]

MedSeq Grounder : [210,157,230,184]

Med Chain R1 (Ours)

<think>The image is a fundus photograph, likely a blue-channel or red-free filter given the monochromatic appearance, which enhances the contrast of the nerve fiber layer and vasculature. The key landmark is the optic disc, the bright, circular area where the retinal vessels converge. Within the optic disc, there is a central, brighter depression known as the optic cup. The region of interest is located centrally within this optic disc. It appears as a distinct, hyper-reflective zone compared to the surrounding neuroretinal rim. Is this the cup or just a reflection artifact? The brightness is consistent with the physiological cupping where the retinal nerve fibers are absent, exposing the underlying lamina cribrosa, which reflects light more strongly. The surrounding darker ring is the neuroretinal rim. Therefore, the central bright area within the disc corresponds to the optic cup.</think> <answer>(213,162),(227,185)</answer>

Figure 25: **Case Study 21: Referring Grounding (Fundus Photography).** Given the referring expression *OpticCup*, the task is to localize the described region. **Qwen3.5** predicts `[600,350,800,550]`. **MedSeq-Grounder** predicts `[210,157,230,184]`. **MEDCHAIN-R1 (Ours)** first locates the optic disc where the retinal vessels converge, then the central bright depression within it, and predicts `(213,162),(227,185)`.