# MedVIGOR: Visual Evidence Internalization for Observation-Driven Reasoning in Medical VLMs

**Yuan Wu**[1,*] **Jiayu Qian**[2,*,✉] **Sipeng Wu**[3,*] **Songpan Gao**[4,5]
**Ruiming Ye**[1] **Zongxian Yang**[1] **Jinyue Li**[6] **Qiankun Li**[2] **Yang Liu**[2]

[1]City University of Hong Kong (Dongguan) [2]Nanyang Technological University
[3]The University of Hong Kong [4]Harbin Institute of Technology, Shenzhen
[5]Pengcheng Laboratory, Shenzhen [6]University of Science and Technology of China

## Abstract

Medical image interpretation requires diagnoses grounded in case-specific visual evidence. However, medical vision-language models often produce plausible answers by exploiting clinical language priors and report-level co-occurrence patterns rather than faithfully using the input image, leading to Visual De-anchoring and Spatial-Semantic Binding Breakdown, where reasoning detaches from visual evidence or binds correct answers to incorrect anatomical support. In this work, we introduce MedVIGOR, a framework that internalizes visual evidence as an intrinsic constraint on medical VLM reasoning. MedVIGOR combines visual-necessity supervision, dual-pathway visual anchor internalization with MedSAM-derived spatial support and DINOv2-derived semantic patterns, and evidence-consistent reinforcement learning to preserve alignment among visual evidence, intermediate reasoning, and final answers. We further present MedVIGOR-Bench, a unified benchmark for evaluating whether medical VLMs are both diagnostically correct and evidentially grounded across concept recognition and visual localization tasks. Experiments show that MedVIGOR improves diagnostic accuracy and grounding fidelity over strong medical VLM baselines, highlighting the value of internalized visual evidence for trustworthy medical reasoning.

## 1 Introduction

Vision-language models (VLMs) have shown strong performance in visual understanding and reasoning [1, 2, 32], motivating growing efforts to adapt them to medical imaging tasks such as diagnosis and report generation [8, 27, 62]. Medical imaging, however, departs sharply from the natural image regime: clinically important findings are often subtle, spatially localized, and embedded in observations with very low signal-to-noise ratio, rather than appearing as salient visual patterns. Reliable interpretation therefore hinges on extracting fine-grained evidence from weak and ambiguous signals.

Under such weak-evidence conditions, medical VLMs can exhibit a characteristic failure mode: rather than relying on case-specific visual evidence, they fall back on medical knowledge priors and report-language co-occurrence patterns acquired during pretraining [18, 35, 54, 67]. We refer to this phenomenon as ***Visual De-anchoring*** **(VDA)**, in which the reasoning process detaches from the input image and drifts toward prior-driven, "closed-book" inference. The risk is amplified in medicine, where highly structured reporting styles and strong co-occurrence patterns among clinical findings make shortcut learning particularly rewarding [33, 56]. Crucially, this failure often remains invisible in the output: the model may still produce fluent, coherent, and partially correct diagnostic statements,

[*]Equal contribution.
[✉]Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

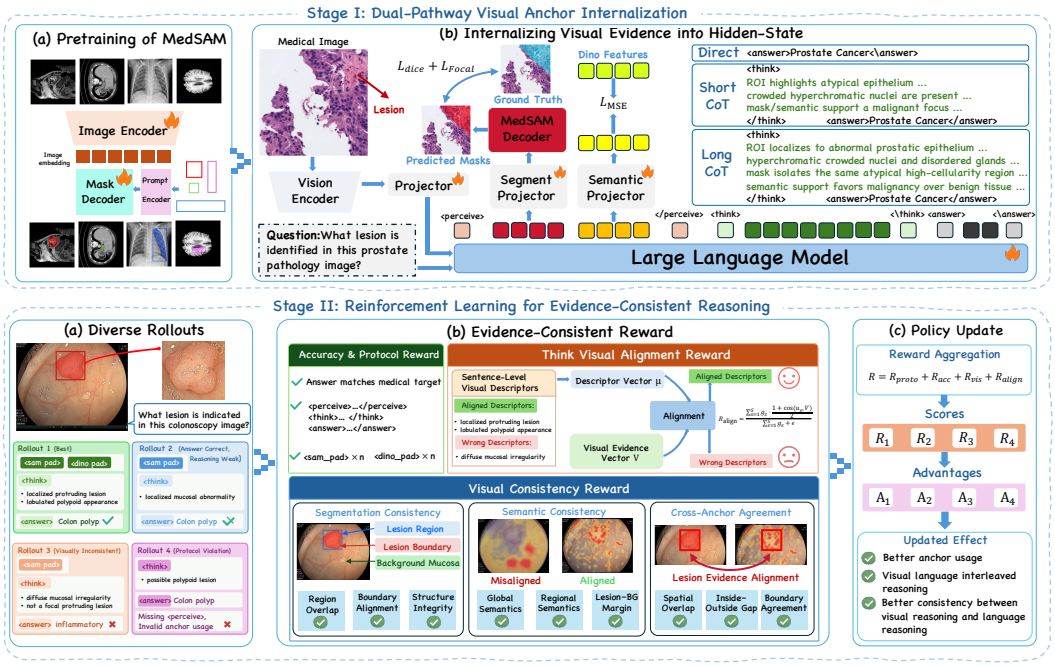



Figure 1: **Overview of MedVIGOR.** MedVIGOR internalizes visual evidence as reasoning constraints through two stages. Stage I supervises anchor states to encode MedSAM-derived spatial support and DINOv2-derived semantic patterns, strengthening spatial-semantic binding. Stage II applies evidence-consistent reinforcement learning to reward protocol-valid, diagnostically correct, visually consistent, and evidence-aligned reasoning.

masking a fundamental gap between *diagnostic plausibility* and *evidential grounding* [14, 25, 29]. The central question for medical VLMs is therefore not whether a model can produce a diagnosis, but whether that diagnosis is grounded in the visual evidence of the case at hand.

Yet evidential grounding is not binary. In clinical reasoning, a diagnosis is trustworthy only when its semantic conclusion is tied to the correct anatomical evidence [11]. This exposes a deeper failure mode: even when a medical VLM yields a correct diagnosis, it may attend to irrelevant regions or mislocalize the pathology, binding the right answer to the wrong evidence [36, 38]. We characterize this structural mismatch between semantic predictions and their supporting spatial evidence as ***Spatial-Semantic Binding Breakdown*** **(SSBB)**. This distinction is clinically consequential: answer correctness and evidential correctness can diverge, and predictions untraceable to a proper anatomical basis are difficult to verify, explain, or trust. Without stable spatial-semantic binding, even superficially correct reasoning falls short of the evidential rigor required in clinical practice.

To address these failures, we propose **Med**ical **V**isual evidence **I**nternalization for **G**rounded **O**bservation-driven **R**easoning (**MedVIGOR**), a framework turning weak external visual evidence into an internal constraint on the reasoning process, as illustrated in Figure 1. MedVIGOR consists of three coordinated components. First, *visual-necessity supervision* constructs training signals whose answers and reasoning traces require case-specific image evidence, reducing reliance on language priors and mitigating VDA. Second, *dual-pathway visual anchor internalization* encodes both spatial support and semantic patterns into hidden states, promoting stable spatial-semantic binding throughout reasoning. Third, *reinforcement learning for evidence-consistent reasoning* preserves this coupling during generation, discouraging drift toward fluent but weakly grounded continuations. Rather than post hoc grounding, MedVIGOR makes visual evidence intrinsic to the reasoning process itself, enabling more reliable medical reasoning under weak-evidence conditions.

To evaluate grounded medical reasoning, we further introduce **MedVIGOR-Bench**, a unified benchmark covering modality, disease and abnormality, anatomy, and visual grounding under both closed- and open-ended protocols. Experiments on MedVIGOR-Bench and standard medical VQA benchmarks show that MedVIGOR improves both diagnostic accuracy and spatial-semantic grounding over strong medical VLM baselines. Our main contributions are as follows:

- **MedVIGOR Framework.** We propose a training paradigm that mitigates visual de-anchoring and spatial-semantic binding breakdown by incorporating visual evidence as an intrinsic constraint within the reasoning hidden states.
- **MedVIGOR-Bench.** We introduce a benchmark for grounded medical reasoning that unifies modality recognition, anatomical identification, and disease diagnosis to evaluate spatial-semantic binding, a dimension often overlooked by standard metrics.
- **Empirical validation.** MedVIGOR attains substantial gains on both accuracy and grounding fidelity; notably, anchor hidden states remain decodable into valid masks and semantic features *even after* external visual extractors are removed at inference.

## 2 Related Works

**Medical Vision-Language Models and Shortcut Learning.** Building on general-domain multimodal foundations [1], recent work adapts vision-language models to medical tasks. Representative systems include LLaVA-Med [27], Med-Gemini [63], and others [12, 39, 53, 59, 64]. While these models achieve strong benchmark performance, such results can mask a weaker form of understanding. Medical images may contain several plausible abnormalities or anatomical structures, so a correct answer does not necessarily identify the observation that supports it. Aggregate accuracy can therefore conflate case-specific recognition with predictions driven by question wording or recurring clinical associations. This distinction is difficult to recover from a fluent explanation alone, especially when the decisive finding is subtle or spatially localized. Recent probing and robustness studies show that VLMs often exploit linguistic priors and co-occurrence patterns rather than case-specific visual content. This behavior is observed in both general-domain settings [35] and in medical studies examining language bias, counterfactual consistency, and guessing-versus-knowing behavior [18, 54, 67]. Prior work largely treats this issue as an input-output bias to be mitigated through additional data or contrastive objectives; we instead frame it as a representational issue that persists across supervision, latent reasoning, and decoding despite apparently correct answers.

**Visual Grounding and Evidence Alignment for Medical VLMs.** Grounding medical model outputs to specific image regions has been approached through several complementary strategies. Phrase-level grounding aligns reports or answers with localized visual evidence [6, 11]. Region-aware generators further condition each sentence on detected anatomical regions [5, 40, 52]. Other approaches incorporate post-hoc fact-checking [16, 36], knowledge-enhanced pretraining [58], or cross-modal disease-clue augmentation [33, 56]. Adjacent work re-accesses visual evidence during generation via visual chain-of-thought [44, 47, 49, 61], or mitigates hallucination at decoding time through contrastive and evidence-aware strategies [14, 25, 29]. Related challenges in grounding and compositionality also persist outside the medical domain [38]. Across these designs, visual evidence is introduced at the input, re-accessed during decoding, or validated after generation. These designs also differ in how explicitly a clinical claim is linked to a particular image region. Region-aware generation can supply localized context [7], whereas decoding-time adjustments and post-hoc checks offer less direct insight into which region informed each intermediate reasoning step. Even when the final answer can be associated with a plausible region, that association does not establish that the region guided the model's reasoning. However, visual evidence is rarely preserved within the intermediate reasoning representation itself. In contrast, our anchor states encode spatial and semantic support directly within the reasoning stream. Existing medical VQA and grounding benchmarks [24, 31] primarily evaluate answer correctness rather than whether predictions are grounded in case-specific evidence, a gap that MedVIGOR-Bench addresses through complementary grounding tasks.

## 3 Methods

### 3.1 Visual-Necessity Supervision

If a training sample can be solved from medical co-occurrence priors or textual leakage alone, the model has little incentive to extract evidence from the image. We therefore construct supervision under a *visual necessity* principle: both the answer and its reasoning trace must depend on information that is only available from the image.

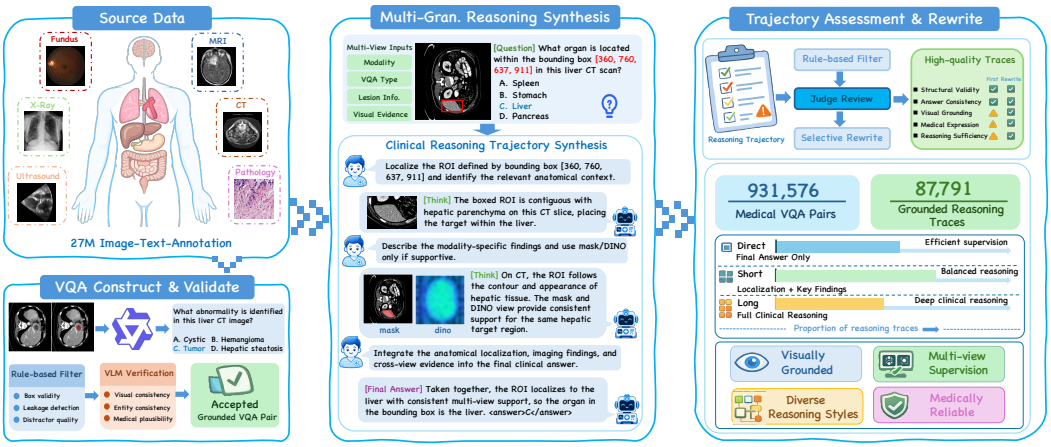



Figure 2: **Visual-necessity supervision construction.** Starting from large-scale medical image–text–annotation data, we construct and validate grounded VQA pairs, synthesize multi-granularity reasoning traces with original, mask, and DINO-based views, and assess trajectories through rule-based filtering, judge review, and selective rewrite. The resulting supervision contains visually grounded, multi-view, and reasoning-style-diverse traces for internalizing medical visual evidence.

As shown in Figure 2, we operationalize this principle through a supervision construction pipeline combining grounded VQA construction and validation, multi-view reasoning trace synthesis, and trajectory assessment with selective rewrite. This pipeline is designed to make both the final answer and intermediate reasoning traces visually grounded rather than solvable from language priors alone.

**Medical VQA Corpus.** We build the corpus from UniBiomed, which contains over 27 million image-text-annotation triplets across 10 biomedical imaging modalities. For each annotated instance, we extract the pixel-level bounding box from the target mask and use the category label together with the original semantic description as visual supervision. Samples are organized into modality recognition, visual grounding, and concept recognition, where concept recognition is further divided into disease/abnormality identification and anatomy identification. Each question-answer pair is constructed to be answerable using image-visible evidence only. To reduce shortcut learning, we impose three constraints: (i) *low-overlap distractor boxes* for grounding, which increase the need for region discrimination; (ii) *box-conditioned anatomy identification*, which forces local visual inspection within a given region; and (iii) *modality recognition without textual leakage*, which removes phrasing-based shortcuts. Rule-based filtering and VLM-based verification further remove invalid boxes, information leakage, hallucinated targets, inconsistent entities, and implausible distractors. This process yields 931,576 medical VQA pairs.

**Multi-View Reasoning Traces.** To make reasoning supervision itself visually dependent, we synthesize traces under multi-view conditioning. Each sample is paired with three complementary views, including the original image, a segmentation mask view, and a DINO-based semantic view. Based on the imaging modality and question type, each sample is further assigned a reasoning style that can be direct, short, or long. The generated traces are organized around target localization, modality-relevant findings, cross-view evidence, and final medical judgment, so that the reasoning chain depends on visual observations from specific views rather than generic clinical priors. We then filter and, when necessary, rewrite traces using rule-based checks and judge-based evaluation over structural validity, answer consistency, visual grounding, medical expression quality, and reasoning sufficiency. This yields 87,791 high-quality reasoning traces. All generation, judging, and rewriting stages are conducted with Qwen3.5-122B-A10B. We further conducted a sampling-based clinical expert audit on 10% of the constructed question–answer pairs and reasoning trajectories. Expert feedback was used for iterative refinement, further improving visual correctness, medical plausibility, and evidence consistency of the final supervision set.

### 3.2 Dual-Pathway Visual Anchor Internalization

Visual-necessity supervision reduces shortcut learning, but it does not ensure that intermediate reasoning states retain image evidence once generation begins. In autoregressive decoding, visual support encoded at the input stage can weaken as the reasoning sequence grows. To mitigate this, we

introduce visual anchor tokens into the reasoning stream and directly supervise their hidden states to encode clinically relevant evidence throughout generation.

**Anchor Design.** Medical reasoning requires both correct localization and correct characterization. We therefore model each anchor with two complementary pathways. A *spatial pathway* captures where the evidence lies using supervision derived from a medical-adapted MedSAM model, preserving lesion or anatomical support and boundary structure. A *semantic pathway* captures what pattern the evidence exhibits using DINOv2 region features, preserving appearance-level semantics beyond explicit localization. This dual-pathway design targets two common errors in medical reasoning: incorrect localization and incorrect morphological interpretation.

**Hidden-State Internalization.** Concretely, we insert a small set of visual anchor tokens into the reasoning sequence. Let $h_a$ denote the hidden state of an anchor token. We supervise $h_a$ through two heads. A lightweight segmentation decoder maps $h_a$ to a spatial prediction trained with $\mathcal{L}_{\text{spatial}}$, and a projection head maps $h_a$ to DINO feature space trained with $\mathcal{L}_{\text{semantic}}$. Together with the response generation loss, the full objective is:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_s \mathcal{L}_{\text{spatial}} + \lambda_d \mathcal{L}_{\text{semantic}}, \tag{1}$$

where $\mathcal{L}_{\text{CE}}$ is the response generation loss, $\mathcal{L}_{\text{spatial}}$ aligns anchor states with spatial supervision, $\mathcal{L}_{\text{semantic}}$ aligns anchor states with DINO semantic features, and $\lambda_s$, $\lambda_d$ are balancing weights. This objective encourages anchor states to encode enough information to recover both spatial support and semantic pattern information, rather than relying only on external visual features at inference time. In Section 5, we verify this property by decoding trained anchor states into segmentation masks and semantic feature maps even after removing external MedSAM and DINOv2 inputs at inference. Further implementation details are provided in the Appendix A.2.

### 3.3 Reinforcement Learning for Evidence-Consistent Reasoning

Although supervised internalization improves evidence retention in anchor states, autoregressive decoding can still drift toward high-probability but weakly grounded continuations. We therefore introduce a reinforcement learning stage to encourage evidence-consistent reasoning during generation.

**Stabilization Rewards.** We first use two stabilization rewards. $R_{\text{proto}}$ enforces a valid reasoning schema through format correctness and anchor-usage consistency. $R_{\text{acc}}$ rewards agreement between the final answer and the ground-truth medical target.

**Visual Consistency Reward.** To maintain evidence fidelity in anchor states during RL, we define:

$$R_{\text{vis}} = \lambda_{\text{seg}} r_{\text{seg}} + \lambda_{\text{dino}} r_{\text{dino}} + \lambda_{\text{cross}} r_{\text{cross}}, \tag{2}$$

where $r_{\text{seg}}$ aligns anchors with spatial supervision, $r_{\text{dino}}$ aligns them with DINO semantic features, and $r_{\text{cross}}$ encourages agreement between the two pathways.

**Reasoning-Evidence Alignment Reward.** Even when anchor states remain faithful, the generated text may still deviate from them. We therefore align sentence-level reasoning with summarized internal visual evidence. Let $V$ denote the summarized anchor representation, and let $\{u_s\}_{s=1}^{S}$ denote sentence-level visual descriptors extracted from the generated reasoning. We define:

$$R_{\text{align}} = \frac{\sum_{s=1}^{S} g_s \cdot \frac{1+\cos(u_s, V)}{2}}{\sum_{s=1}^{S} g_s + \epsilon}, \tag{3}$$

where $g_s$ is the importance weight of the $s$-th sentence.

**Optimization.** The final reward is:

$$R = R_{\text{proto}} + R_{\text{acc}} + R_{\text{vis}} + R_{\text{align}}, \tag{4}$$

where visual rewards are applied only when the generated response satisfies the required protocol. We optimize the policy with GRPO:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q,\{o_i\}_{i=1}^{G}} \left[ \frac{1}{G} \sum_{i=1}^{G} \min(\rho_i A_i,\ \text{clip}(\rho_i, 1-\epsilon, 1+\epsilon) A_i) \right], \tag{5}$$

where $\rho_i = \frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}$ is the policy ratio and $A_i$ is the advantage computed from the aggregated reward $R$. Further implementation details are provided in the Appendix A.3.

# 4 MedVIGOR-Bench

## 4.1 Why MedVIGOR-Bench?

Existing medical VLM benchmarks primarily assess final-answer correctness, which does not establish whether a prediction is supported by the relevant visual evidence. This distinction is particularly important in medical imaging, where a plausible response may arise from learned clinical regularities rather than observations specific to the image at hand. A model may identify the correct finding without locating its supporting region, or attend to a relevant region while assigning it the wrong medical meaning. Answer-level metrics provide limited insight into these failures, leaving the relationship among concept recognition, anatomical interpretation, and evidence localization insufficiently examined. Evaluating grounded reasoning therefore requires tasks that probe whether medical concepts can be associated with the specific visual content that supports them, in addition to measuring answer correctness. We introduce **MedVIGOR-Bench** as a unified framework for assessing both medical concept recognition and its spatial association with image evidence.

## 4.2 Benchmark Design and Protocol

MedVIGOR-Bench comprises 13,000 questions, including 8,938 closed-ended and 4,062 open-ended questions. It covers four complementary task families. *Modality Recognition* evaluates the identification of imaging modalities, while *Disease and Abnormality* assesses the recognition of image-visible pathological findings. *Anatomy Identification* requires associating a specified bounding box with the corresponding anatomical structure. Conversely, *Visual Grounding* requires selecting, from candidate regions, the one that supports a given medical concept.

The closed-ended protocol covers all four task families, whereas the open-ended protocol covers modality recognition, disease and abnormality, and anatomy identification through free-form responses. CT and MRI are evaluated as annotated two-dimensional slices under the same single-image protocol used for other modalities. We report per-task accuracy, sample-weighted averages within each protocol, and an overall accuracy aggregated across all evaluated questions. Representative examples and detailed task statistics are provided in Appendix B.

# 5 Experiments

## 5.1 Experimental Setup

**Implementation details.** We implement MedVIGOR based on Qwen2.5-VL-7B [3]. All training experiments run on 8 NVIDIA H100 GPUs. In the dual-pathway visual anchor internalization stage, we train on 87,791 visually grounded reasoning samples, including direct, short-form, and long-form reasoning traces. The model is optimized for 2 epochs with a learning rate of $5 \times 10^{-5}$. During subsequent RL stage, we use 90,000 samples and train for 2 epochs with a learning rate of $1 \times 10^{-6}$. Unless otherwise specified, all evaluation results are obtained from the final MedVIGOR checkpoint.

**Baseline methods.** We compare MedVIGOR with four categories of baselines: open-source general VLMs from the Qwen-VL [2–4] and InternVL [9, 10, 55] families; medical specialist VLMs such as MedVLM-R1 [41], HuatuoGPT-Vision [8], Lingshu [62], and Hulu-Med [19]; tool-augmented reasoning models such as DeepEyes [65], Mini-o3 [22], and PixelReasoner-RL [51]; and multimodal medical agents such as MMedAgent [26], MedAgent-Pro [57], and MDAgents [21]. All models are evaluated using the same prompts, answer extraction rules, and scoring protocols.

**Evaluation metrics.** We adopt accuracy as the unified metric across all MedVIGOR-Bench tasks. For closed-ended questions, predictions are scored by exact match after answer extraction. For visual grounding, a prediction is correct if the selected region matches the ground-truth supporting region among the candidate boxes. For open-ended questions, we use GPT-4o as an LLM judge to assess clinical equivalence against the reference answer under a fixed rubric and decoding configuration. We report task-level accuracy and an aggregated overall score over all evaluated samples. All experimental results are averaged over multiple runs.

Table 1: **Results on MedVIGOR-Bench.** Accuracy (%) is reported under closed-ended and open-ended. **Avg.** denotes the sample-weighted average within each protocol, while **Overall** aggregates all evaluated samples across both protocols. Best results are **bold**, and second-best are underlined.

| Method | Closed-ended | | | | | Open-ended | | | | Overall |
|---|---|---|---|---|---|---|---|---|---|---|
| | Mod. Rec. | Dis. & Abn. | Anat. ID | Visual Ground. | **Avg.** | Mod. Rec. | Dis. & Abn. | Anat. ID | **Avg.** | |
| *Frontier closed-source VLMs* | | | | | | | | | | |
| GPT-5 | 96.91 | 48.32 | 56.37 | 74.63 | 70.85 | 95.38 | 37.25 | 40.79 | 56.72 | 66.44 |
| Gemini2.5-Pro [13] | 96.47 | 72.45 | 65.75 | 69.51 | 73.77 | 86.07 | 33.83 | 22.48 | 47.71 | 65.63 |
| *Open-source general VLMs* | | | | | | | | | | |
| InternVL3-8B [55] | 99.12 | 53.45 | 52.91 | 51.09 | 59.06 | 97.38 | 30.46 | 27.43 | 51.11 | 56.58 |
| Qwen3-VL-8B [4] | 97.87 | 48.72 | 55.76 | 65.85 | 66.26 | 95.46 | 23.96 | 30.59 | 48.47 | 60.70 |
| Qwen3.5-9B | 97.65 | 66.63 | 52.91 | 62.61 | 67.67 | 89.99 | 27.38 | 26.44 | 47.17 | 61.26 |
| *Tool-augmented VLMs* | | | | | | | | | | |
| DeepEyes-7B [65] | 96.40 | 41.47 | 34.76 | 58.13 | 58.12 | 92.07 | 10.27 | 27.92 | 40.82 | 52.72 |
| Mini-o3-7B [22] | 85.66 | 26.93 | 11.42 | 40.21 | 41.36 | 91.38 | 10.61 | 25.15 | 40.05 | 40.95 |
| PixelReasoner-7B [51] | 94.12 | 44.55 | 49.64 | 64.43 | 63.43 | 89.53 | 5.19 | 24.95 | 37.08 | 55.19 |
| *Medical specialist VLMs* | | | | | | | | | | |
| MedVLM-R1-2B [41] | 96.84 | 46.55 | 36.29 | 29.62 | 43.90 | 87.53 | 36.51 | 17.13 | 48.01 | 45.18 |
| HuatuoGPT-Vision [8] | 99.41 | 45.75 | 37.51 | 62.49 | 62.08 | 98.00 | 35.14 | 31.68 | 54.38 | 59.68 |
| Lingshu-7B [62] | 99.04 | 65.43 | 59.63 | 58.96 | 66.40 | 98.38 | 40.90 | 33.66 | 57.48 | 63.62 |
| Hulu-Med [19] | **99.56** | 71.76 | 61.67 | 36.66 | 55.86 | 97.15 | 44.44 | 27.33 | 57.04 | 56.23 |
| *Multimodal medical agents* | | | | | | | | | | |
| MMedAgent-7B [26] | 95.37 | 45.52 | 49.75 | 56.52 | 59.53 | 92.53 | 36.39 | 26.93 | 51.99 | 57.18 |
| MedAgent-Pro [57] | 96.99 | 39.19 | 35.78 | 59.41 | 58.57 | 93.15 | 5.02 | 27.35 | 38.75 | 52.38 |
| MDAgents [21] | 96.03 | 47.12 | 56.07 | 68.29 | 67.02 | 93.38 | 25.73 | 34.06 | 49.43 | 61.52 |
| Qwen2.5-VL-7B [3] | 97.06 | 42.21 | 49.13 | 53.12 | 57.23 | 94.53 | 16.83 | 28.81 | 44.66 | 53.30 |
| + Answer-direct SFT | 98.53 | 82.32 | 73.80 | 65.21 | 74.58 | 97.54 | 52.99 | 41.98 | 64.50 | 71.43 |
| **MedVIGOR** | 99.49 | **94.30** | **90.01** | **89.20** | **91.86** | **98.69** | **79.46** | **49.01** | **78.04** | **87.54** |

## 5.2 Main Results

**Results on MedVIGOR-Bench.** Table 1 shows that MedVIGOR obtains the highest closed-ended average (91.86%), open-ended average (78.04%), and overall accuracy (87.54%). The answer-direct SFT baseline uses the same number of supervised training examples and improves the overall accuracy of the Qwen2.5-VL-7B backbone from 53.30% to 71.43%. MedVIGOR further improves on this baseline by 17.28 percentage points in the closed-ended average, 13.54 points in the open-ended average, and 16.11 points overall.

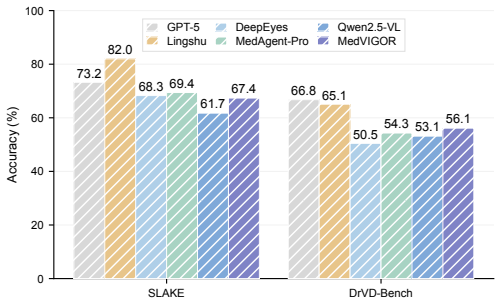



Figure 3: **External benchmark check.**

At the task level, the largest gaps are in open-ended Disease and Abnormality (79.46% compared with 52.99%) and closed-ended Visual Grounding (89.20% compared with 65.21%). Visual Grounding accounts for more than half of the closed-ended questions, so this gain has a large effect on the closed-ended average. MedVIGOR also raises closed-ended Anatomy Identification from 73.80% to 90.01% and open-ended Anatomy Identification from 41.98% to 49.01%. Both models already perform well on Modality Recognition, where the gains are smaller. Compared with the unadapted backbone, MedVIGOR improves the closed-ended average from 57.23% to 91.86% and the open-ended average

from 44.66% to 78.04%. We next examine whether these gains reflect stronger alignment between predictions and their supporting visual evidence.

**Results on external benchmarks.** Figure 3 reports external benchmark checks on SLAKE [31] and DrVD-Bench [66]. These datasets are not used as the primary evidence for our grounded-reasoning claim; instead, they examine whether MedVIGOR preserves general medical VQA ability beyond MedVIGOR-Bench. MedVIGOR improves over its Qwen2.5-VL backbone on both datasets, suggesting that the proposed training does not merely fit the constructed benchmark. At the same time, some medical specialist models may remain stronger on standard medical VQA datasets, likely due to domain-specific pretraining or instruction tuning on similar data distributions. We therefore treat these external results as complementary transfer evidence, while focusing the main analysis on evidence-grounded reasoning in MedVIGOR-Bench.

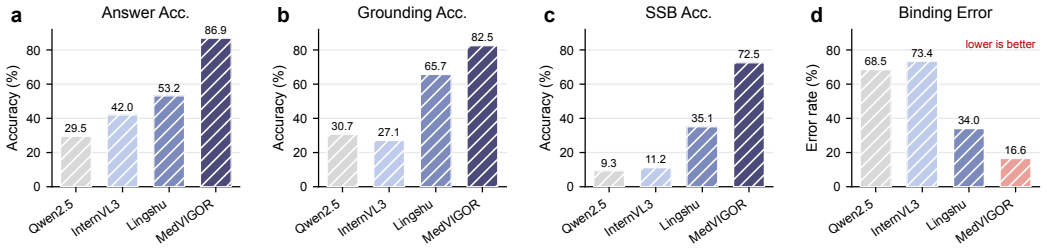



Figure 4: **Evidence grounding analysis on Disease and Abnormality samples.** MedVIGOR improves answer accuracy, grounding accuracy, and SSB accuracy while reducing BER, showing that its diagnostic predictions are more consistently coupled with the correct supporting regions.

### 5.3 Evidence Grounding Analysis

We examine whether correct diagnostic answers are grounded in case-specific visual evidence. Let $A$ denote a correct semantic answer and $G$ denote correct grounding. We define spatial-semantic binding accuracy as $\mathrm{SSBAcc} = P(A \cap G)$ and binding error rate as $\mathrm{BER} = P(\neg G \mid A)$.

On 3,506 Disease and Abnormality samples with paired answer and grounding labels, Figure 4 further shows that MedVIGOR improves answer accuracy from 29.5% to 86.9% and grounding accuracy from 30.7% to 82.5% compared with the Qwen2.5-VL-7B backbone. More importantly, it raises SSB Accuracy from 9.3% to 72.5% while reducing Binding Error Rate from 68.5% to 16.6%, showing stronger coupling between diagnostic predictions and supporting regions. Figure 5 evaluates evidence sensitivity through text-only, shuffled-image, target-region-masked, and non-target-region-masked settings. MedVIGOR degrades substantially under text-only and shuffled-image inputs, indicating reduced reliance on language priors or mismatched visual context. It also shows a larger drop when the target evidence region is masked than when non-target regions are masked, suggesting that its predictions selectively depend on diagnostically relevant evidence. This provides complementary perturbation-level evidence for stronger spatial-semantic binding.

Table 2 further tests whether this coupling depends on internalized anchor evidence. We intervene on visual-anchor states at inference time while keeping model parameters and non-anchor reasoning states unchanged. Removing, shuffling, zeroing, or noise-corrupting these states consistently reduces answer accuracy, grounding accuracy, and SSB Accuracy, while increasing Binding Error Rate. This indicates that visual-anchor states are functional carriers of spatial-semantic evidence rather than auxiliary

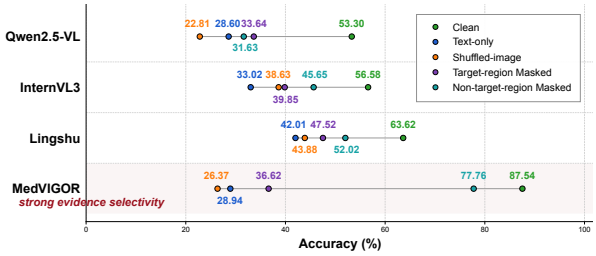



Figure 5: **Visual perturbation sensitivity.**

Table 2: **Anchor causal interventions.**

| Intervention | Ans. | Gnd. | SSB | BER↓ |
|---|---|---|---|---|
| **Full MedVIGOR** | **86.88** | **82.49** | **72.50** | **16.55** |
| Remove tokens | 75.22 | 60.46 | 43.90 | 41.64 |
| Shuffle states | 69.43 | 47.51 | 28.27 | 59.28 |
| Zero states | 71.14 | 53.67 | 35.41 | 50.22 |
| Noise states | 66.33 | 44.28 | 22.86 | 65.54 |

Table 3: **Two-stage ablations of MedVIGOR.** VG denotes visual grounding accuracy. Blue parentheses report absolute gains: Stage I variants are compared with VNS Only, except VNS Only is compared with Qwen2.5-VL; Stage II variants are compared with Stage I Full.

**(a) Stage I: Visual Anchor Internalization**

| Method | VNS | Spatial | Semantic | Overall | VG |
|---|---|---|---|---|---|
| Qwen2.5-VL | – | – | – | 53.30 | 53.12 |
| VNS Only | ✓ | – | – | 77.83 (+24.53) | 74.62 (+21.50) |
| Spatial Only | ✓ | ✓ | – | 80.76 (+2.93) | 84.26 (+9.64) |
| Semantic Only | ✓ | – | ✓ | 80.12 (+2.29) | 83.99 (+9.37) |
| **Stage-I Full** | ✓ | ✓ | ✓ | **82.82** (+4.99) | **85.98** (+11.36) |

**(b) Stage II: Evidence-Consistent RL**

| Method | $R_{\text{proto}}$ | $R_{\text{acc}}$ | $R_{\text{vis}}$ | $R_{\text{align}}$ | Overall | VG |
|---|---|---|---|---|---|---|
| w/o $R_{\text{proto}}$ | – | ✓ | ✓ | ✓ | 85.46 (+2.64) | 86.92 (+0.94) |
| w/o $R_{\text{acc}}$ | ✓ | – | ✓ | ✓ | 83.71 (+0.89) | 86.76 (+0.78) |
| w/o $R_{\text{vis}}$ | ✓ | ✓ | – | ✓ | 86.13 (+3.31) | 86.15 (+0.17) |
| w/o $R_{\text{align}}$ | ✓ | ✓ | ✓ | – | 86.94 (+4.12) | 87.46 (+1.48) |
| **MedVIGOR** | ✓ | ✓ | ✓ | ✓ | **87.54** (+4.72) | **89.20** (+3.22) |

formatting tokens. For example, noise corruption reduces SSB Accuracy from $72.50\%$ to $22.86\%$ and increases Binding Error Rate from $16.55\%$ to $65.54\%$. These results provide causal evidence that MedVIGOR's grounding fidelity is supported by its internalized visual-anchor representations.

### 5.4 Ablation Studies

We conduct ablation studies to examine the contribution of each training component in MedVIGOR. As shown in Table 3, VNS Only already improves the overall score from $53.30\%$ to $77.83\%$ and visual grounding accuracy from $53.12\%$ to $74.62\%$, showing the importance of visually necessary supervision. Adding either Spatial or Semantic anchor supervision further improves performance, while Stage I Full achieves the best Stage I results, indicating that spatial support and semantic visual patterns provide complementary evidence. Stage II Evidence-Consistent RL brings additional gains over Stage I Full, improving the overall score from $82.82\%$ to $87.54\%$ and visual grounding accuracy from $85.98\%$ to $89.20\%$. The reward ablations show that each component contributes to the final model: removing $R_{\text{acc}}$ causes the largest overall drop, while removing $R_{\text{vis}}$ most clearly affects visual grounding accuracy. $R_{\text{proto}}$ and $R_{\text{align}}$ further stabilize valid anchor usage and evidence-aligned reasoning. Figure 6 analyzes the ablation experiment of SAM and DINO anchors. The balanced 4 SAM anchors and 4 DINO anchors setting achieves near-saturated performance, whereas increasing the number of anchors yields limited additional gains while incurring extra computational cost. Together, these results show that MedVIGOR benefits from the combination of visually necessary supervision, dual-pathway anchor supervision, and evidence-consistent reward optimization.

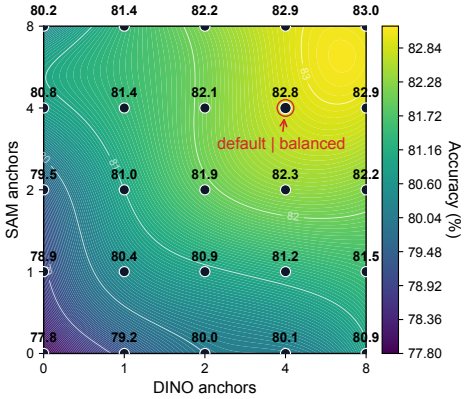



Figure 6: **Anchor configuration ablation.**

### 5.5 Case Study

Figure 7 shows an anatomy-identification example in cardiac ultrasound. HuatuoGPT-Vision predicts the wrong answer, aortic valve, because its reasoning relies on generic anatomical priors rather than the boxed visual evidence, reflecting Visual De-anchoring. Lingshu produces the correct answer, left ventricle, but its explanation is only loosely tied to the target region and still refers to valve-like echoes, indicating Spatial-Semantic Binding Breakdown.

In contrast, MedVIGOR treats the bounding box as the primary evidence region and identifies a large hypoechoic chamber bordered by myocardial tissue, which is more consistent with ventricular anatomy than with a compact valve or tubular vessel. Rather than relying on the broad location of the structure, its reasoning uses chamber size, echogenicity, and surrounding tissue to distinguish the left ventricle from the other answer choices. The anatomical conclusion thus follows an explicit description of the local visual evidence. The decoded anchor mask also aligns with the target region, suggesting that the internalized anchor states preserve spatial evidence relevant to the final prediction. This comparison illustrates that answer correctness alone is insufficient: a model may be wrong due to prior-driven reasoning, or correct for weakly localized reasons. This case shows

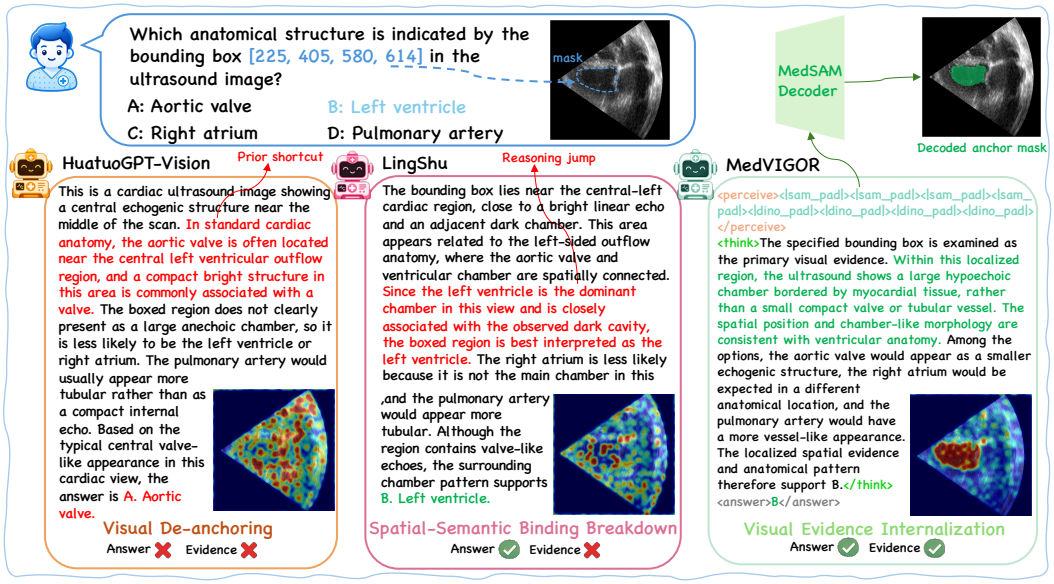



Figure 7: **Case study on evidence-consistent anatomy identification in cardiac ultrasound.** MedVIGOR correctly identifies the left ventricle while maintaining alignment between the predicted answer, localized visual evidence, and decoded anchor mask, whereas baseline models exhibit visual de-anchoring or spatial-semantic binding breakdown.

that MedVIGOR improves not only answer correctness, but also the coupling between semantic prediction and localized visual evidence.

## 6 Conclusion

In this work, we investigated grounded reasoning in medical vision-language models and identified two key failure modes: Visual De-anchoring, where models rely on clinical priors rather than case-specific image evidence, and Spatial-Semantic Binding Breakdown, where correct predictions are not properly tied to their supporting anatomical regions. To address these challenges, we proposed MedVIGOR, a framework that internalizes visual evidence into the reasoning process through visual-necessity supervision, dual-pathway visual anchor internalization, and evidence-consistent reinforcement learning. Experiments on MedVIGOR-Bench demonstrate that MedVIGOR improves both diagnostic accuracy and visual grounding over strong general, medical-specialist, tool-augmented, and agent-based baselines. Further perturbation analyses and anchor-state interventions show that the learned visual anchors actively support the coupling between semantic predictions and localized evidence. Overall, MedVIGOR highlights the importance of moving beyond final-answer accuracy toward medical multimodal systems whose reasoning remains faithful to case-specific visual evidence, providing a step toward more accurate, grounded, and verifiable medical VLMs.

## Acknowledgements

This work was supported by the National Key R&D Program of China under Grant No. 2025YFE0216500; by the National Research Foundation Singapore and the Cyber Security Agency under the National Cyber-security R&D Programme (NCRP25-P04-TAICeN); and by the Prime Minister's Office, Singapore under the Campus for Research Excellence and Technological Enterprise (CREATE) Programme. This research is also supported by AI Innovation and Application Joint Research Institute, Shenzhen Technology University (Grants No. 20261064010117). Any opinions, findings and conclusions, or recommendations expressed in these materials are those of the author(s) and do not reflect the views of the National Research Foundation, Singapore, Cyber Security Agency of Singapore, Singapore. The authors declare no competing interests.

## References


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

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL https://arxiv.org/abs/2308.12966.

[3] Shuai Bai et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[4] Shuai Bai et al. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

[5] 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.

[6] 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*, pages 1–21. Springer, 2022.

[7] Aofei Chang, Le Huang, Alex James Boyd, Parminder Bhatia, Taha Kass-Hout, Fenglong Ma, and Cao Xiao. Medsight: Towards grounded visual comprehension in medical large vision-language models. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

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

[9] 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*, pages 24185–24198, 2024.

[10] Zhe Chen et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

[11] Zhihao Chen, Yang Zhou, Anh Tran, Junting Zhao, Liang Wan, Gideon Su Kai Ooi, Lionel Tim-Ee Cheng, Choon Hua Thng, Xinxing Xu, Yong Liu, et al. Medical phrase grounding with region-phrase context contrastive alignment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 371–381. Springer, 2023.

[12] 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*, 2024.

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

[14] Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14303–14312, 2024.

[15] 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.

[16] Alice Heiman, Xiaoman Zhang, Emma Chen, Sung Eun Kim, and Pranav Rajpurkar. Factchexcker: Mitigating measurement hallucinations in chest x-ray report generation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30787–30796, 2025.

[17] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Xu Tang, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.

[18] Songtao Jiang, Yuxi Chen, Sibo Song, Yan Zhang, Yeying Jin, Yang Feng, Jian Wu, and Zuozhu Liu. Knowing or guessing? robust medical visual question answering via joint consistency and contrastive learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 325–335. Springer, 2025.

[19] Songtao Jiang et al. Hulu-Med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*, 2025.

[20] Wentao Jiang, Xiang Feng, Zengmao Wang, Yong Luo, Pingbo Xu, Zhe Chen, Bo Du, and Jing Zhang. Rex-rag: Reasoning exploration with policy correction in retrieval-augmented generation. *arXiv preprint arXiv:2508.08149*, 2025.

[21] Yubin Kim et al. MDAgents: An adaptive collaboration of LLMs for medical decision-making. *arXiv preprint arXiv:2404.15155*, 2024.

[22] Xin Lai, Junyi Li, Wei Li, Tao Liu, Tianjian Li, and Hengshuang Zhao. Mini-o3: Scaling up reasoning patterns and interaction turns for visual search. *arXiv preprint arXiv:2509.07969*, 2025.

[23] 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.

[24] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1): 180251, 2018.

[25] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882, 2024.

[26] Binxu Li, Tiankai Yan, Yuanting Pan, Jie Luo, Ruiyang Ji, Jiayuan Ding, Zhe Xu, Shilong Liu, Haoyu Dong, Zihao Lin, et al. Mmedagent: Learning to use medical tools with multi-modal agent. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8745–8760, 2024.

[27] 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.

[28] Jinyue Li, Yuci Liang, Qiankun Li, Xinheng Lyu, Jiayu Qian, Huabao Chen, Kun Wang, Zhigang Zeng, Anil Anthony Bharath, and Yang Liu. Pathmem: Toward cognition-aligned memory transformation for pathology mllms, 2026. URL https://arxiv.org/abs/2603.09943.

[29] Zehui Liao, Shishuai Hu, Ke Zou, Huazhu Fu, Liangli Zhen, and Yong Xia. Vision-amplified semantic entropy for hallucination detection in medical visual question answering. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 669–679. Springer, 2025.

[30] Tianwei Lin et al. HealthGPT: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. *arXiv preprint arXiv:2502.09838*, 2025.

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

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

[33] Kang Liu, Zhuoqi Ma, Xiaolu Kang, Zhusi Zhong, Zhicheng Jiao, Grayson Baird, Harrison Bai, and Qiguang Miao. Structural entities extraction and patient indications incorporation for chest x-ray report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 433–443. Springer, 2024.

[34] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2034–2044, 2025.

[35] Tiange Luo, Ang Cao, Gunhee Lee, Justin Johnson, and Honglak Lee. Probing visual language priors in vlms. *arXiv preprint arXiv:2501.00569*, 2024.

[36] Razi Mahmood, Diego Machado-Reyes, Joy Wu, Parisa Kaviani, Ken CL Wong, Niharika D'Souza, Mannudeep Kalra, Ge Wang, Pingkun Yan, and Tanveer Syeda-Mahmood. Phrase-grounded fact-checking for automatically generated chest x-ray reports. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 441–452. Springer, 2025.

[37] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning, 2025. URL https://arxiv.org/abs/2503.07365.

[38] Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14420–14431, 2024.

[39] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine learning for health (ML4H)*, pages 353–367. PMLR, 2023.

[40] Philip Müller, Georgios Kaissis, and Daniel Rueckert. Chex: Interactive localization and region description in chest x-rays. In *European Conference on Computer Vision*, pages 92–111. Springer, 2024.

[41] Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 337–347. Springer, 2025.

[42] Liangrui Pan, Jiadi Luo, Chenchen Nie, Songqing Fan, Xiang Wang, Shaoliang Peng, and Qingchun Liang. A deep learning framework integrating tumor microenvironmental features accurately predicts multiple driver gene mutations in lung cancer pathology images. *Cancer Research*, 86(5):1319–1337, 2026.

[43] Liangrui Pan, Jiadi Luo, Yuxuan Xiao, Chenchen Nie, Xiaoshuai Wu, Songqing Fan, Ling Chu, Manqiu Li, Rongfang He, Zhenyu Zhao, et al. Diffusion attention expert model for predicting and semi-automatic localizing stas in lung cancer histopathological images. *Nature Communications*, 17:6905, 2026.

[44] Ji Qi, Ming Ding, Weihan Wang, Yushi Bai, Qingsong Lv, Wenyi Hong, Bin Xu, Lei Hou, Juanzi Li, Yuxiao Dong, et al. Cogcom: Train large vision-language models diving into details through chain of manipulations. 2024.

[45] Jiayu Qian, Yuwu Lu, Wuyuan Xie, Zhihui Lai, Miaohui Wang, and Xuelong Li. Cpsr-clip: Conditional prompt-induced style reconstruction for zero-shot domain adaptation. *IEEE Transactions on Multimedia*, 27:7714–7728, 2025. doi: 10.1109/TMM.2025.3599044.

[46] Jiayu Qian, Zongxian Yang, Guanxing Chen, Pengwei Hu, KC Tan, Yan Wang, Yu-An Huang, and Zhi-An Huang. Hyperbolic relational prompts for intersectional fairness in medical vlms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13712–13721, June 2026.

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

[48] 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.

[49] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024.

[50] 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.

[51] Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*, 2025.

[52] Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7433–7442, 2023.

[53] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138, 2024.

[54] Xingyu Wan, Qiaoying Teng, Jun Chen, Yonghan Lu, Deqi Yuan, and Zhe Liu. Eliminating language bias for medical visual question answering with counterfactual contrastive training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 194–204. Springer, 2025.

[55] Weiyun Wang et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

[56] Yuhao Wang, Yue Sun, Tao Tan, Chao Hao, Yawen Cui, Xinqi Su, Weichen Xie, Linlin Shen, and Zitong Yu. Trrg: Towards truthful radiology report generation with cross-modal disease clue enhanced large language models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 647–657. Springer, 2025.

[57] Ziyue Wang, Junde Wu, Chang Han Low, and Yueming Jin. MedAgent-Pro: Towards multi-modal evidence-based medical diagnosis via reasoning agentic workflow. *arXiv preprint arXiv:2503.18968*, 2025.

[58] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 21372–21383, 2023.

[59] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 16(1):7866, 2025.

[60] Sipeng Wu, Shiqin Tang, Chenxi Liu, Jingwei Wu, Wenjing Jiang, Xiaoyu Xiao, Hongbin Liu, and Gaofeng Meng. TRACE-PSG: Traceable Sleep Diagnosis Support with Pointer-Grounded Evidence . In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2026*, volume LNCS 16886. Springer Nature Switzerland, September 2026.

[61] Yuan Wu, Zongxian Yang, Jiayu Qian, Songpan Gao, Guanxing Chen, Qiankun Li, Yu-An Huang, and Zhi-An Huang. Better eyes, better thoughts: Why vision chain-of-thought fails in medicine. *arXiv preprint arXiv:2603.06665*, 2026.

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

[63] Lin Yang, Shawn Xu, Andrew Sellergren, Timo Kohlberger, Yuchen Zhou, Ira Ktena, Atilla Kiraly, Faruk Ahmed, Farhad Hormozdiari, Tiam Jaroensri, et al. Advancing multimodal medical capabilities of gemini. *arXiv preprint arXiv:2405.03162*, 2024.

[64] Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D Davison, Hui Ren, et al. A generalist vision–language foundation model for diverse biomedical tasks. *Nature medicine*, 30(11):3129–3141, 2024.

[65] Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. DeepEyes: Incentivizing "thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025.

[66] Tianhong Zhou, Yingtao Zhu, Chuxi Xiao, Haiyang Bian, Lei Wei, Xuegong Zhang, et al. Drvd-bench: Do vision-language models reason like human doctors in medical image diagnosis? *Advances in Neural Information Processing Systems*, 38, 2026.

[67] Huanjia Zhu, Yishu Liu, Chengju Zhou, Guangming Lu, and Bingzhi Chen. Med-biasx: Robust medical visual question answering with language biases. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 369–378. Springer, 2025.

## Appendix Contents


**A Additional Method Details** **17**

A.1 Visual-Necessity Supervision . . . . . 17

A.2 Dual-Pathway Visual Anchor Internalization . . . . . 17

A.3 Reinforcement Learning for Evidence-Consistent Reasoning . . . . . 18

A.4 The Algorithm Workflow of MedVIGOR . . . . . 20

**B MedVIGOR-Bench** **20**

**C Additional Related works** **21**

C.1 Reinforcement Learning for Multimodal Reasoning . . . . . 21

**D Training Configuration** **21**

**E Qualitative Examples and Failure Cases** **22**

E.1 Qualitative Examples . . . . . 22

E.2 Failure Cases . . . . . 23

**F Prompts** **23**

F.1 Prompts for Data and Reasoning Supervision . . . . . 23

F.2 Prompts for Model Protocol and Evaluation . . . . . 24

**G Limitations** **24**

# A Additional Method Details

## A.1 Visual-Necessity Supervision

We provide additional implementation details for the visual-necessity data pipeline. After converting each valid target mask into a tight absolute-coordinate bounding box, we apply lightweight structural checks before model-based verification. The rule-based stage removes samples with invalid or out-of-bound boxes, degenerate masks, boxes covering more than 60% of the image, leaked multiple-choice prefixes, missing box conditions for anatomy questions, or non-visual clinical cues such as patient history, symptoms, demographic information, and laboratory findings. The remaining samples are judged by Qwen3.5-122B-A10B under five criteria: consistency with the annotated target, absence of unsupported visual claims, explicit entity presence for grounding questions, absence of answer leakage for identification and modality questions, and plausibility of distractors for multiple-choice questions.

Figure 8 summarizes the composition of the constructed supervision. Figure 8(a) shows broad modality coverage, with MRI and CT as the dominant sources and X-ray, ultrasound, and other modalities providing additional diversity. Figures 8(b) and 8(c) show that the anatomical coverage includes both high-frequency organ systems and long-tail structures, rather than being concentrated on a small set of common regions. Figure 8(d) further shows that the supervision contains anatomical entities, pathological concepts, and modality-related descriptors, supporting the mixture of anatomy identification, abnormality recognition, and grounding-oriented samples.

Reasoning styles are assigned according to question type and expected visual reasoning complexity. Modality recognition and bounding-box selection are generally routed to the direct style, since their supervision target is compact. Local anatomy and simple region-conditioned questions are routed to short rationales. Diagnostic, attribute-based, pathology-related, CT/MRI interpretation, and option-aware differential questions are routed to long rationales, where explicit links between visible findings and the final medical judgment are required. The routed assignments are rebalanced to avoid over-concentration in a single style. As shown in Figure 8(e), the final trace set contains 39,626 short traces, 26,304 direct traces, and 21,861 long traces.

Generated traces are normalized and checked for tag validity, answer consistency, length constraints, modality-specific terminology, cross-modality conflicts, generic filler, meta-instruction leakage, repetition, and missing differential reasoning when applicable. Non-direct traces are further judged for visual grounding, medical specificity, and reasoning sufficiency. If either rule-based checks or the judge indicate insufficient quality, the trace is rewritten from scratch using the same visual inputs and then rejudged before finalization. This rewrite–rejudge loop reduces noisy rationales while preserving the original answer supervision.

## A.2 Dual-Pathway Visual Anchor Internalization

**Shared pathway design.** Both the spatial and semantic pathways share an identical lightweight adapter: hidden states at the pathway-specific anchor positions (<|sam_pad|> and <|dino_pad|> in the <perceive> span) are passed through a LayerNorm and a learned linear projection, after which a small set of learned queries cross-attends to the projected states to produce pathway outputs $O^s$ and $O^d$. The spatial pathway additionally applies rotary position encoding before projection to preserve 2-D layout. The two pathways differ only in the target space of the projection and the supervision signal applied to the output.

**Spatial supervision.** The spatial output $O^s$ is fed to the frozen MedSAM decoder as prompt embeddings, together with the MedSAM image embedding $E^s$ obtained by one forward pass of the MedSAM image encoder on the input image: $\hat{M} = \mathcal{D}_{\mathrm{SAM}}(E^s, O^s)$. Predicted anchor masks are matched to ground-truth masks by Hungarian assignment with a Dice cost, and supervised using a combined Dice and BCE loss on matched pairs, plus an empty-mask penalty on unmatched ones:

$$\mathcal{L}_{\mathrm{spatial}} = \frac{1}{|\pi^*|} \sum_{(i,j)\in\pi^*} \left[1 - \mathrm{Dice}(\hat{M}_i, G_j) + \mathrm{BCE}(\hat{M}_i, G_j)\right] + \gamma \frac{1}{|\mathcal{U}|} \sum_{k\in\mathcal{U}} \mathrm{BCE}(\hat{M}_k, \mathbf{0}), \qquad (6)$$

where $\pi^*$ is the matched assignment and $\mathcal{U}$ is the set of unmatched anchors. Samples without mask annotations contribute only the background term.

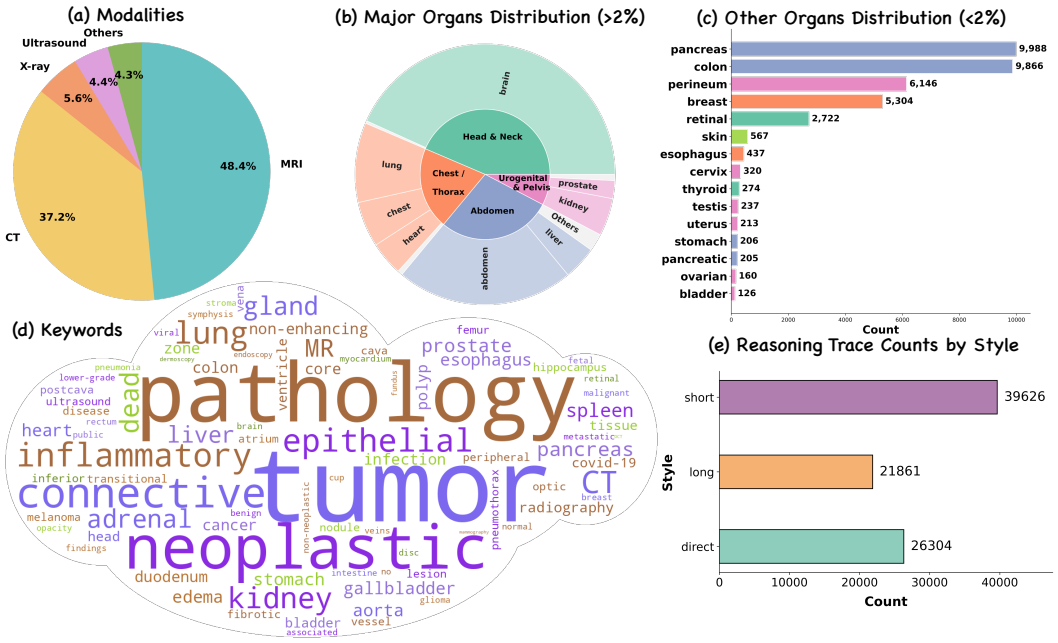



Figure 8: Data composition of visual-necessity supervision. The figure reports modality coverage, frequent and long-tail anatomical distributions, keyword statistics, and the final reasoning-style distribution.

**Semantic supervision.** The semantic output $O^d$ is distilled from the frozen DINOv2 teacher features $F^d$ of the same image via MSE loss, $\mathcal{L}_{\text{semantic}} = \frac{1}{N_d}\|O^d - F^d\|_F^2$, where $N_d$ is the number of semantic tokens. This preserves region-level semantic identity without committing to pixel-accurate boundaries, and is complementary to the pixel-level signal of the spatial pathway.

**Training objective.** The full internalization loss combines language modeling with the two evidence-grounding terms:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_s \mathcal{L}_{\text{spatial}} + \lambda_d \mathcal{L}_{\text{semantic}}. \tag{7}$$

MedSAM and DINOv2 remain frozen throughout training; only the anchor projections, cross-attention queries, and the base VLM are optimized.

### A.3 Reinforcement Learning for Evidence-Consistent Reasoning

This appendix completes the reward specification in Section 3.3. We make three things explicit that the main text only sketches: (i) the components inside the stabilization rewards $R_{\text{proto}}$ and $R_{\text{acc}}$; (ii) how the pathway-specific scores $r_{\text{seg}}, r_{\text{dino}}, r_{\text{cross}}$ are computed and combined into $R_{\text{vis}}$; and (iii) how the summarized evidence $V$ is instantiated as a three-axis descriptor $V_i$ and paired with sentence-level descriptors $u_s$ inside $R_{\text{align}}$.

**Stabilization rewards.** For the $i$-th rollout, $R_{\text{proto},i}$ averages a format score and an anchor-usage score,

$$R_{\text{proto},i} = \tfrac{1}{2}\left(r_i^{\text{fmt}} + r_i^{\text{perc}}\right), \tag{8}$$

where $r_i^{\text{fmt}} \in [0,1]$ is the fraction of required tags that appear in the correct order (`<perceive>` $\prec$ `<think>` $\prec$ `<answer>`, each opened and closed exactly once), and $r_i^{\text{perc}} \in [0,1]$ rewards the `<perceive>` span for containing the task-expected number and types of anchor tokens, decaying linearly with count mismatch. The accuracy reward $R_{\text{acc},i}$ is task-dependent: exact match for closed-form answers, and a normalized token-overlap score against the medical target for open-ended ones.

**Visual gate.** Before the three pathway-specific visual rewards are aggregated, each is multiplied by a shared gate that suppresses contributions from malformed rollouts:

$$g_i^{\text{vis}} = \text{clip}\left(\frac{r_i^{\text{fmt}}}{\tau_{\text{fmt}}}, 0, 1\right) \cdot \text{clip}(r_i^{\text{perc}}, 0, 1), \qquad \tau_{\text{fmt}} = 0.60, \tag{9}$$

where $r_i^{\text{fmt}}$ and $r_i^{\text{perc}}$ are the format and anchor-count scores from Eq. 8. This is the precise form of the protocol-score gating referenced in Section 3.3: visual signals are linearly attenuated when format compliance falls below $\tau_{\text{fmt}}$ or the anchor count is wrong, and pass through unchanged otherwise.

**Visual Consistency Reward.** The three pathways probe complementary aspects of evidence preservation: *seg* scores the spatial fidelity of decoded masks, *dino* scores the semantic fidelity of decoded features, and *cross* scores the coherence between the two. Each pathway-specific score is a gated, weighted sum of sub-scores:

$$r_{\text{raw}}^k = \sum_j \alpha_j^k \, s_j^k, \qquad r_i^k = \text{clip}\big(g_i^{\text{vis}} \, r_{\text{raw}}^k, \, 0, \, 1\big), \qquad k \in \{\text{seg, dino, cross}\}. \tag{10}$$

All sub-scores $s_j^k \in [0, 1]$ are evaluated between rollout-time anchor decodings and frozen teacher signals (MedSAM masks and DINOv2 features, consistent with Section 3.2); cosine-based sub-scores are mapped to $[0, 1]$ via $(1 + \cos)/2$ before weighting, while Dice, IoU, and count-agreement scores are natively non-negative. Definitions and weights $\alpha_j^k$ are listed in Table 4. The aggregated visual consistency reward is

$$R_{\text{vis},i} = \lambda_{\text{seg}} \, r_i^{\text{seg}} + \lambda_{\text{dino}} \, r_i^{\text{dino}} + \lambda_{\text{cross}} \, r_i^{\text{cross}}, \tag{11}$$

which closes the loop with the $\lambda$-weighted form of Section 3.3; values of $\lambda_{\text{seg}}, \lambda_{\text{dino}}, \lambda_{\text{cross}}$ are reported in Appendix D.

Table 4: Sub-scores composing the three pathway-specific visual rewards. All sub-scores are normalized to $[0, 1]$; the weights $\alpha_j^k$ sum to one within each pathway.

| Pathway | Sub-score | Definition | $\alpha$ |
|---|---|---|---|
| seg | $s_{\text{region}}^{\text{seg}}$ | Hungarian-matched mean Dice with bg-coverage offset and unmatched penalty | 0.45 |
| | $s_{\text{boundary}}^{\text{seg}}$ | Mean boundary IoU over matched mask pairs | 0.35 |
| | $s_{\text{struct}}^{\text{seg}}$ | Connected-component and hole count agreement vs. GT | 0.20 |
| dino | $s_{\text{global}}^{\text{dino}}$ | Weighted cosine of [CLS] and mean-patch features vs. teacher | 0.20 |
| | $s_{\text{region}}^{\text{dino}}$ | Mean cosine over lesion/boundary/bg region-pooled features | 0.30 |
| | $s_{\text{relation}}^{\text{dino}}$ | Cosine of upper-triangular Gram vectors from normalized patch features | 0.30 |
| | $s_{\text{margin}}^{\text{dino}}$ | Lesion-vs-bg separation margin against teacher prototypes | 0.20 |
| cross | $s_{\text{overlap}}^{\text{cross}}$ | Soft-IoU plus cosine of SAM lesion map and DINO saliency map | 0.45 |
| | $s_{\text{margin}}^{\text{cross}}$ | Inside/outside DINO similarity margin vs. lesion/bg prototypes | 0.35 |
| | $s_{\text{boundary}}^{\text{cross}}$ | Cosine of SAM-boundary-pooled DINO feature and teacher boundary prototype | 0.20 |

**Reasoning-Evidence Alignment Reward.** To compute $R_{\text{align},i}$, we first instantiate the summarized evidence $V$ from Section 3.3 as a three-dimensional descriptor $V_i$:

$$\begin{aligned} V_i = \Big( & \underbrace{\tfrac{1}{2}\big(s_{\text{overlap}}^{\text{cross}} + s_{\text{margin}}^{\text{cross}}\big)}_{\text{spatial agreement}}, \\ & \underbrace{\tfrac{1}{3}\big(s_{\text{boundary}}^{\text{seg}} + s_{\text{struct}}^{\text{seg}} + s_{\text{boundary}}^{\text{cross}}\big)}_{\text{boundary/structure fidelity}}, \\ & \underbrace{\tfrac{1}{3}\big(s_{\text{region}}^{\text{dino}} + s_{\text{relation}}^{\text{dino}} + s_{\text{margin}}^{\text{dino}}\big)}_{\text{regional semantics}} \Big). \end{aligned} \tag{12}$$

The three coordinates recast $V_i$ into the same three-axis space we use for per-sentence descriptors—*where* the evidence is, *how it is bounded*, and *what it depicts*—the aspects most directly verifiable from a textual statement. For each sentence $s = 1, \ldots, S$ in the `<think>` span, a lightweight text-side parser produces a scalar importance weight $g_s \in [0, 1]$ and a descriptor $u_s \in [0, 1]^3$ in the same space; sentences that carry no visual claim receive $g_s = 0$ and drop out of the sum. The alignment reward is

$$R_{\text{align},i} = \frac{\sum_{s=1}^{S} g_s \cdot \frac{1+\cos(u_s, V_i)}{2}}{\sum_{s=1}^{S} g_s + \epsilon_0}, \tag{13}$$

**Algorithm 1 MedVIGOR**: Two-Stage Algorithm Workflow

**Require:** Base VLM $\pi_{\theta_0}$, CoVT data $\mathcal{D}_c$, RL data $\mathcal{D}_r$, visual anchors $\{\text{MedSAM}, \text{DINOv2}\}$, reward weights $\omega$
**Ensure:** Optimized policy $\pi_\theta$
Initialize $\pi_\theta \leftarrow \pi_{\theta_0}$ and add special tokens $\langle perceive\rangle, \langle think\rangle, \langle answer\rangle$.
**Stage I: Dual-Pathway Visual Anchor Internalization**
**for** each batch $(I, Q, A, M) \sim \mathcal{D}_c$ **do**
    Construct structured target:

$$y = \langle perceive\rangle z^{sam}, z^{dino} \langle /perceive\rangle \langle think\rangle r \langle /think\rangle \langle answer\rangle A \langle /answer\rangle.$$

    Forward $\pi_\theta$ and extract anchor hidden states $h^{sam}$ and $h^{dino}$.
    Decode anchor states into spatial masks and semantic features.
    Optimize

$$\mathcal{L}_{\text{int}} = \mathcal{L}_{\text{CE}} + \lambda_s \mathcal{L}_{\text{spatial}} + \lambda_d \mathcal{L}_{\text{semantic}}.$$

**end for**
**Stage II: Reinforcement Learning for Evidence-Consistent Reasoning**
**for** each prompt $(I, Q) \sim \mathcal{D}_r$ **do**
    Generate $G$ rollouts $\{y_g\}_{g=1}^{G}$ from $\pi_\theta$.
    Compute protocol and answer rewards for each rollout.
    Decode rollout-time anchor states and compute visual consistency and reasoning-evidence alignment rewards.
    Aggregate rewards:

$$R_g = \sum_k \omega_k r_{g,k}.$$

    Compute group-relative advantages from $\{R_g\}_{g=1}^{G}$ and update $\pi_\theta$ with GRPO.
**end for**
**return** $\pi_\theta$

where $\epsilon_0 = 10^{-6}$ is a numerical floor (distinct from the GRPO clipping radius in Section 3.3). The formula self-regulates two degenerate cases: when the `<think>` span contains no visually grounded sentences, every $g_s = 0$ and the numerator vanishes, driving $R_{\text{align},i} \to 0$ through the $\epsilon_0$-stabilized denominator; when the visual descriptor $V_i$ has negligible norm and the cosine is numerically undefined, we set $R_{\text{align},i} = 0$ by convention.

**Optimization.** The aggregated reward $R_i = R_{\text{proto},i} + R_{\text{acc},i} + R_{\text{vis},i} + R_{\text{align},i}$ feeds the GRPO objective of Section 3.3. Together, the four reward terms ask the policy to improve answer correctness while jointly maintaining protocol validity, spatial grounding, semantic consistency, and alignment between textual reasoning and the internalized visual evidence.

### A.4 The Algorithm Workflow of MedVIGOR

This subsection presents the full training workflow of MedVIGOR. Algorithm 1 outlines the two-stage procedure, including the dual-pathway visual anchor internalization and the reinforcement learning for evidence-consistent reasoning.

## B MedVIGOR-Bench

Figure 9 illustrates representative examples from the four task families of MedVIGOR-Bench: Modality Recognition, Disease and Abnormality Recognition, Anatomy Identification, and Visual Grounding. The benchmark is organized to evaluate distinct but complementary aspects of grounded medical visual reasoning. Modality Recognition measures whether a model can infer the imaging modality from visual evidence. Disease and Abnormality Recognition evaluates the identification of image-visible pathological findings. Anatomy Identification tests region-conditioned anatomical understanding by requiring the model to associate a specified bounding box or visual region with the corresponding anatomical structure. Visual Grounding evaluates evidence localization by asking the model to select the candidate region that best supports a given medical concept. Together, these

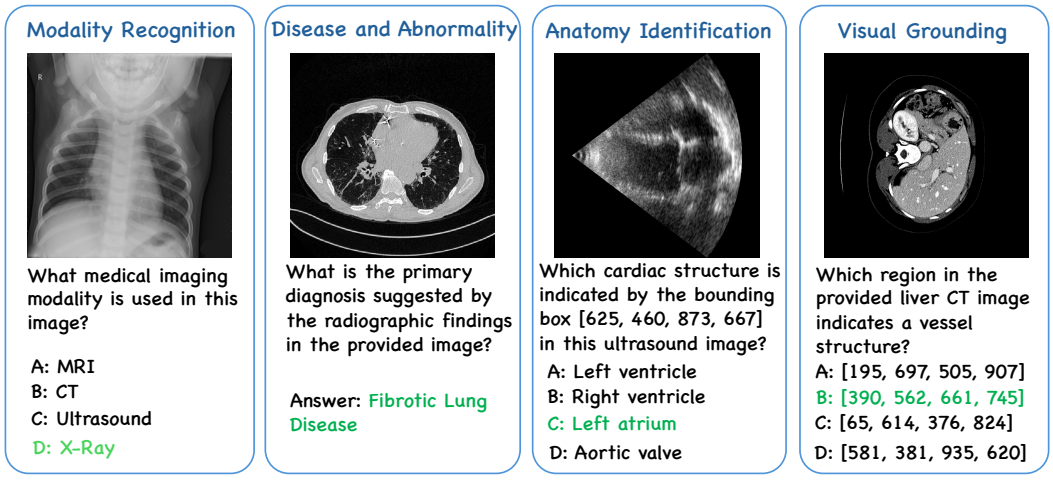



Figure 9: Representative Examples of MedVIGOR-Bench Task Families.

tasks provide a unified protocol for assessing not only semantic recognition, but also the association between medical answers and their supporting visual evidence.

MedVIGOR-Bench contains 13,000 questions, consisting of 8,938 closed-ended questions and 4,062 open-ended questions. The closed-ended split includes all four task families and supports deterministic scoring through answer extraction and exact matching. In particular, the Visual Grounding task is formulated as candidate-region selection, where a prediction is considered correct only when the selected region matches the ground-truth supporting region. The open-ended split includes Modality Recognition, Disease and Abnormality Recognition, and Anatomy Identification, and is intended to evaluate whether models can produce clinically meaningful free-form answers beyond fixed answer choices. Table 5 summarizes the task-level composition of the benchmark. In addition to the main results reported in the paper, Table 6 provides additional model comparisons on MedVIGOR-Bench, complementing the main results reported in Table 1.

# C Additional Related works

## C.1 Reinforcement Learning for Multimodal Reasoning

Recent advances in deep learning and vision-language models have demonstrated substantial progress in visual representation learning, medical image analysis, and clinical reasoning [28, 42, 43, 45, 46]. Building on these advances, reinforcement learning has emerged as a promising paradigm for enhancing structured reasoning, with verifiable rewards in the language domain [15, 50] extending to multimodal settings through GRPO-style optimization with vision-grounded signals [17, 34, 37]. Such approaches have also been explored in medical applications, including medical VQA, report generation, and evidence-grounded diagnosis [41, 60]. In retrieval-augmented language-model reasoning, REX-RAG explores alternative reasoning trajectories and corrects for sampling-induced policy shifts [20]. However, existing multimodal RL methods primarily rely on answer correctness, format compliance, or high-level response-image consistency, leaving the internal reasoning states that produce the answer largely unconstrained. In contrast, we directly reward the evidential content of anchor hidden states and their alignment with generated reasoning text, penalizing language-prior-driven trajectories even when their final answers are correct.

# D Training Configuration

This section reports the reproducibility-oriented training configurations used in MedVIGOR. The settings cover MedSAM pretraining, dual-pathway visual anchor internalization, and reinforcement learning for evidence-consistent reasoning, as summarized in Table 7.

Table 5: Task distribution of MedVIGOR-Bench.

| Protocol | Modality Recognition | Disease & Abnormality | Anatomy Identification | Visual Grounding | Total |
|---|---|---|---|---|---|
| Closed-ended | 1360 | 1753 | 981 | 4844 | 8,938 |
| Open-ended | 1299 | 1753 | 1010 | – | 4,062 |
| Total | 2659 | 3506 | 1991 | 4844 | 13,000 |

Table 6: **Additional model comparisons on MedVIGOR-Bench.** We include representative models from the main table as anchors and report additional baselines evaluated under the same protocol. Accuracy (%) is reported under closed-ended and open-ended protocols. **Avg.** denotes the sample-weighted average within each protocol, while **Overall** aggregates all evaluated samples across both protocols.

| Method | Closed-ended | | | | | Open-ended | | | | Overall |
|---|---|---|---|---|---|---|---|---|---|---|
| | Mod. Rec. | Dis. & Abn. | Anat. ID | Visual Ground. | **Avg.** | Mod. Rec. | Dis. & Abn. | Anat. ID | **Avg.** | |
| GPT-5 | 96.91 | 48.32 | 56.37 | 74.63 | 70.85 | 95.38 | 37.25 | 40.79 | 56.72 | 66.44 |
| Gemini2.5-Pro [13] | 96.47 | 72.45 | 65.75 | 69.51 | 73.77 | 86.07 | 33.83 | 22.48 | 47.71 | 65.63 |
| Gemma-31B | 99.78 | 65.77 | 54.84 | 50.35 | 61.39 | 94.92 | 34.17 | 19.50 | 49.95 | 57.82 |
| Qwen2.5-VL-32B [3] | 96.25 | 49.46 | 59.23 | 69.74 | 68.64 | 93.23 | 20.76 | 35.45 | 47.59 | 62.06 |
| Qwen3-VL-8B [4] | 97.87 | 48.72 | 55.76 | 65.85 | 66.26 | 95.46 | 23.96 | 30.59 | 48.47 | 60.70 |
| Qwen3-VL-30B [4] | **99.93** | 47.80 | 54.13 | 71.33 | 69.18 | 97.31 | 34.91 | 34.65 | 54.80 | 64.68 |
| Qwen3.5-9B | 97.65 | 66.63 | 52.91 | 62.61 | 67.67 | 89.99 | 27.38 | 26.44 | 47.17 | 61.26 |
| Qwen3.5-35B | 97.50 | 69.14 | 60.14 | 66.60 | 71.09 | 81.83 | 32.00 | 18.02 | 44.46 | 62.77 |
| InternVL2.5-8B [10] | 99.26 | 53.74 | 53.01 | 42.98 | 54.75 | 97.84 | 57.50 | 17.33 | 60.41 | 56.52 |
| InternVL2.5-38B [10] | 98.60 | 70.56 | 59.12 | 46.20 | 60.37 | 94.53 | 65.20 | 18.22 | 62.90 | 61.16 |
| MedGemma-4B [48] | 91.47 | 63.09 | 51.27 | 35.69 | 51.26 | 90.69 | 35.42 | 18.02 | 48.77 | 50.48 |
| InternVL3-8B [55] | 99.12 | 53.45 | 52.91 | 51.09 | 59.06 | 97.38 | 30.46 | 27.43 | 51.11 | 56.58 |
| InternVL3-38B [55] | 98.01 | 51.68 | 60.55 | 73.27 | 71.40 | 97.23 | 31.37 | 22.48 | 50.22 | 64.78 |
| LLaVA-Med-7B [27] | 42.57 | 14.15 | 5.61 | 23.53 | 22.62 | 88.76 | 7.19 | 15.64 | 35.38 | 26.61 |
| MedVLM-R1-2B [41] | 96.84 | 46.55 | 36.29 | 29.62 | 43.90 | 87.53 | 36.51 | 17.13 | 48.01 | 45.18 |
| Med-R1-2B [23] | 95.96 | 43.75 | 33.44 | 31.19 | 43.76 | 85.60 | 43.53 | 17.92 | 50.62 | 45.90 |
| HuatuoGPT-Vision-7B [8] | 99.41 | 45.75 | 37.51 | 62.49 | 62.08 | 98.00 | 35.14 | 31.68 | 54.38 | 59.68 |
| Lingshu-7B [62] | 99.04 | 65.43 | 59.63 | 58.96 | 66.40 | 98.38 | 40.90 | 33.66 | 57.48 | 63.62 |
| Hulu-Med-7B [19] | 99.56 | 71.76 | 61.67 | 36.66 | 55.86 | 97.15 | 44.44 | 27.33 | 57.04 | 56.23 |
| HealthGPT-Pro-8B [30] | 99.19 | 51.28 | 55.35 | 41.27 | 53.59 | 98.08 | 37.31 | 27.43 | 54.28 | 53.81 |
| Qwen2.5-VL-7B [3] | 97.06 | 42.21 | 49.13 | 53.12 | 57.23 | 94.53 | 16.83 | 28.81 | 44.66 | 53.30 |
| **MedVIGOR-7B** | 99.49 | **94.30** | **90.01** | **89.20** | **91.86** | **98.69** | **79.46** | **49.01** | **78.04** | **87.54** |

All MedVIGOR training experiments are conducted on 8 NVIDIA H100 GPUs using mixed-precision training. We use DeepSpeed ZeRO-2 for multi-GPU optimization. The MedSAM pretraining, Stage I dual-pathway visual anchor internalization, and Stage II reinforcement learning stages take approximately 72, 6, and 48 hours. In addition to model training, the construction of reasoning supervision uses Qwen3.5-122B-A10B for generation, judging, and rewriting, while open-ended evaluation uses GPT-4o under a fixed rubric.

# E Qualitative Examples and Failure Cases

## E.1 Qualitative Examples

Figures 10–12 show representative qualitative examples. Compared with baseline models, MedVIGOR more reliably links the final answer to the corresponding visual evidence, especially for localization, abnormality recognition, and region-conditioned anatomy identification.

Table 7: Training configurations of MedVIGOR. The table summarizes the settings used for MedSAM pretraining, Stage I dual-pathway visual anchor internalization, and Stage II reinforcement learning with evidence-consistent rewards.

| Stage | Configuration | Value |
|---|---|---|
| *Global setup* | | |
| | Base VLM | Qwen2.5-VL-7B |
| | Training hardware | 8 NVIDIA H100 GPUs |
| *MedSAM pretraining* | | |
| | Backbone | MedSAM-ViT-B |
| | Optimizer | AdamW |
| | Learning rate | $1 \times 10^{-4}$ |
| | Weight decay | 0.01 |
| | Global batch size | 64 |
| | Epochs | 150 |
| | Segmentation loss | Dice loss + BCE loss |
| *Stage I: Dual-pathway visual anchor internalization* | | |
| | Optimizer | AdamW |
| | Learning rate | $5 \times 10^{-5}$ |
| | Weight decay | 0.1 |
| | Warmup ratio | 0.05 |
| | Global batch size | 96 |
| | Epochs | 2 |
| | LoRA rank / alpha | 16 / 32 |
| | Spatial loss weight $\lambda_s$ | 1.0 |
| | Semantic loss weight $\lambda_d$ | 1.0 |
| *Stage II: Evidence-consistent reinforcement learning* | | |
| | Optimizer | AdamW |
| | Learning rate | $1 \times 10^{-6}$ |
| | Global batch size | 48 |
| | Epochs | 2 |
| | LoRA rank / alpha | 16 / 32 |
| | Rollout group size | 4 |
| | Segmentation reward weight $\lambda_{\text{seg}}$ | 0.15 |
| | DINO reward weight $\lambda_{\text{dino}}$ | 0.10 |
| | Cross-anchor reward weight $\lambda_{\text{cross}}$ | 0.10 |

### E.2 Failure Cases

Figures 13 and 14 show representative failure cases. These examples suggest that further improvements are needed for challenging cases involving subtle visual evidence and fine-grained evidence binding.

## F Prompts

This section provides the prompts used in data construction, reasoning supervision, and evaluation. These prompts define the visual-only constraints, response format, filtering criteria, and open-ended answer judging protocol used throughout the experiments.

### F.1 Prompts for Data and Reasoning Supervision

Figures 15–20 show the prompts for VQA construction, verification, reasoning generation, judging, and rewriting. These prompts are used to ensure that questions and rationales depend on image-visible evidence.

### F.2 Prompts for Model Protocol and Evaluation

Figure 21 shows the open-ended evaluation prompt used to judge answer equivalence.

## G Limitations

While MedVIGOR improves evidence-grounded medical reasoning, several limitations remain. First, our experiments are conducted with a 7B-parameter backbone and primarily focus on single-image medical VQA, leaving the effectiveness of the proposed framework across model scales and more complex clinical tasks unexplored. Second, although our evidence-alignment rewards encourage consistency between visual representations and generated reasoning, they do not guarantee that the model causally relies on the internalized visual evidence when producing its final predictions. Third, the training framework relies on external visual supervision and annotated medical resources to construct evidence-aware rewards, which may limit its applicability to modalities or clinical conditions with scarce annotations. Furthermore, the use of selected anchor hidden states may not fully capture fine-grained or spatially distributed visual evidence, particularly for subtle abnormalities and ambiguous anatomical boundaries. Future work will investigate more direct measures of evidence faithfulness, annotation-efficient reward construction, and extensions to multi-image reasoning and clinical report generation.

## Question

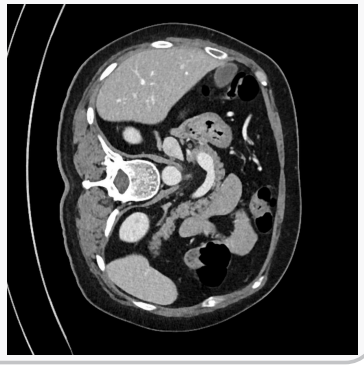
Where is the liver located in this abdominal CT scan?

A: [74, 610, 617, 932]

B: [428, 0, 741, 186]

C: [260, 126, 639, 351]

D: [739, 183, 1015, 347]



## MedVLM

To determine the location of the liver in the abdominal CT scan, we need to identify the liver's position relative to the other structures in the image.

1. **Identify the liver's position in the image:**

-The liver is located in the upper right quadrant of the abdomen.

2. **Analyze the CT scan:**

-The liver is located in the area marked as [74, 610, 617, 932].

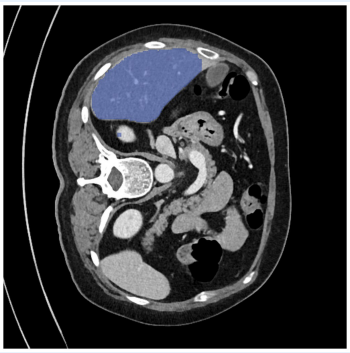
3. **Match the position to the options:**

- The coordinates [74, 610, 617, 932] are in the upper right quadrant of the abdomen, which matches the position of the liver.

Therefore, the correct answer is:A: [74, 610, 617, 932]

## MedVIGOR

Decode



<perceive><|sam_pad|><|sam_pad|><|sam_pad|><|sam_pad|><|dino_pad|><|dino_pad|><|dino_pad|><|dino_pad|><|dino_pad|></perceive>

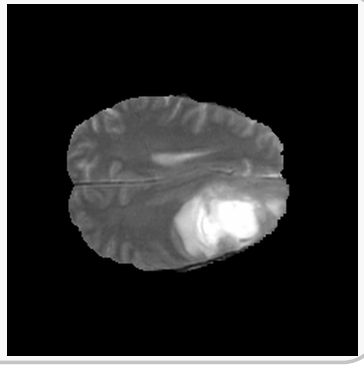
<think>The axial CT image demonstrates a large, homogenous organ occupying the right upper quadrant, consistent with the anatomical position of the liver. The segmentation mask highlights this specific region, confirming the boundaries of the hepatic parenchyma. The organ exhibits uniform soft-tissue attenuation without focal hypodense or hyperdense lesions, and its smooth contour is clearly delineated against adjacent structures like the stomach and spleen. This appearance is characteristic of normal liver tissue, distinguishing it from the other abdominal organs listed in the options. The location and radiographic features align with the coordinates provided in option C, which accurately encompasses the right-sided hepatic mass.</think><answer>C</answer>

Figure 10: Qualitative example on abdominal CT grounding.

**Question**

What pathological structure is visualized in the center of the brain scan?

Answer: tumor

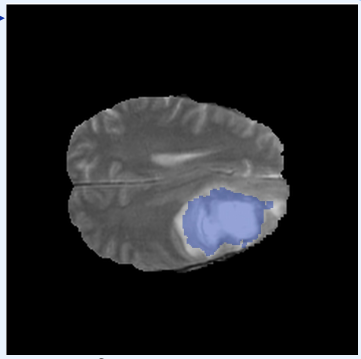


**MedGemma**

Based on the image, the pathological structure visualized in the center of the brain scan appears to be a **large, well-defined lesion**. It is located in the left hemisphere of the brain.
Without further clinical information or additional imaging, it is difficult to determine the exact nature of the lesion. However, based on the appearance, it could potentially be a **hemorrhage, tumor, or other mass lesion**.
It is important to note that this is just a preliminary observation based on the image alone. A definitive diagnosis requires a full clinical evaluation, review of the patient's history, and potentially additional imaging studies.

**MedVIGOR**

Decode



<perceive><|sam_pad|><|sam_pad|><|sam_pad|><|sam_pad|><|dino_pad|><|dino_pad|><|dino_pad|><|dino_pad|></perceive>
The axial T2-weighted MRI slice demonstrates a large, well-circumscribed, hyperintense mass located in the right cerebral hemisphere, specifically involving the parietal and occipital lobes. This lesion exhibits significant mass effect, evidenced by the compression of the adjacent lateral ventricle and a noticeable midline shift toward the left side. The segmentation mask precisely delineates the boundaries of this abnormality, confirming its substantial volume and irregular but defined margins. The DINO semantic view further supports the presence of a distinct pathological entity by highlighting the central core of the lesion with high contrast against the surrounding parenchyma. The hyperintense signal on T2-weighted imaging is characteristic of high water content, which is typical for neoplastic processes such as gliomas or metastases, often associated with peritumoral edema. While acute ischemic stroke can also present with hyperintensity, the mass effect and the specific morphology of the lesion favor a neoplastic etiology rather than a vascular event. The absence of a clear vascular territory distribution and the presence of a discrete, space-occupying lesion make a tumor the most likely diagnosis. Therefore, the visualized structure is identified as a tumor.</think><answer>tumor</answer>

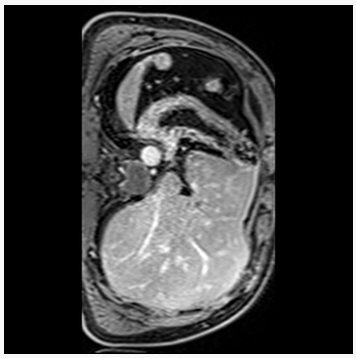
Figure 11: Qualitative example on brain MRI abnormality recognition.

**Question**

Which anatomical structure is highlighted by the bounding box [284, 429, 743, 897] in this abdominal MRI scan?

A: Spleen B: Liver

C: Kidney D: Pancreas



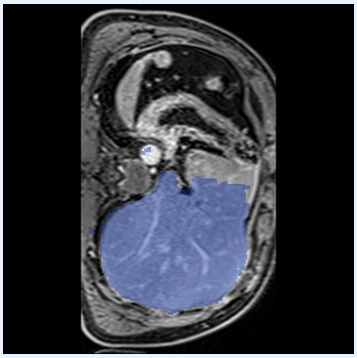
**HuatuoGPT-Vision**

The bounding box [284, 429, 743, 897] in the abdominal MRI scan highlights the **Spleen**.

The spleen is located in the upper left quadrant of the abdomen and is typically visible in MRI scans due to its distinct shape and position relative to other organs.

**MedVIGOR**

Decode



<perceive><|sam_pad|><|sam_pad|><|sam_pad|><|sam_pad|><|dino_pad|><|dino_pad|><|dino_pad|><|dino_pad|></perceive>
<think>The bounding box highlights a large, solid organ occupying the right upper quadrant of the abdomen, consistent with the anatomical location of the liver. On this axial MRI slice, the organ demonstrates relatively homogeneous intermediate signal intensity, distinct from the adjacent hyperintense fluid-filled structures and the spleen located on the contralateral side. The segmentation mask confirms the extensive parenchymal volume and smooth contour typical of hepatic tissue, while the DINO view reinforces the structural integrity of this large abdominal mass. Given the size, position, and signal characteristics relative to surrounding organs, the findings are diagnostic of the liver rather than the spleen, kidney, or pancreas.</think><answer>B</answer>

Figure 12: Qualitative example on abdominal MRI anatomy identification.

**Question**

What anatomical structure is located at the bounding box [193, 431, 323, 589] in this retinal fundus image?
A: Macula lutea
B: Central retinal artery
C: Optic disc D: Vitreous body

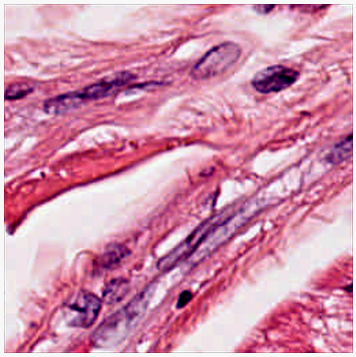


**MedVIGOR**

<perceive><|sam_pad|><|sam_pad|><|sam_pad|><|sam_pad|><|dino_pad|><|dino_pad|><|dino_pad|><|dino_pad|></perceive>
<think>The bounding box [193, 431, 323, 589] localizes the central region of the optic disc, which is the anatomical structure responsible for the convergence of retinal vessels. In the fundus image, this area appears as a bright, circular region with a distinct yellowish hue, corresponding to the macula lutea, situated temporally relative to the optic disc. The segmentation mask confirms the precise boundaries of this central structure, distinguishing it from the surrounding retinal vasculature and the optic disc itself. The DINO view further highlights the central circular region, consistent with the macula lutea's location. Given the bright, circular appearance and central position, the structure is identified as the macula lutea rather than the optic disc or central retinal artery.</think><answer>A</answer>

Figure 13: Failure case on fundus anatomy identification.

**Question**



Which histological finding is observed in the provided bladder tissue specimen?
A: Nuclei-level inflammatory infiltration
B: Epithelial hyperplasia with keratinization
C: Fibrosis with smooth muscle proliferation
D: Glomerular capillary loop thickening

**MedVIGOR**

<perceive><|sam_pad|><|sam_pad|><|sam_pad|><|sam_pad|><|dino_pad|><|dino_pad|><|dino_pad|><|dino_pad|></perceive>
<think>The provided histological image displays a section of bladder tissue characterized by a distinct interface between the epithelial lining and the underlying stroma. The epithelial cells exhibit significant nuclear crowding and pleomorphism, with nuclei appearing enlarged and hyperchromatic, which are hallmark features of cellular atypia. There is a notable absence of a dense inflammatory infiltrate within the stromal compartment, and the connective tissue does not show signs of extensive fibrosis or smooth muscle proliferation. The epithelial cells display a stratified arrangement with some cells appearing keratinized, indicating a disruption of normal polarity and a potential pre-neoplastic or neoplastic process. The DINO semantic view highlights the heterogeneity of the cellular regions, reinforcing the presence of abnormal cellular clusters rather than uniform stromal tissue. This pattern of epithelial hyperplasia with keratinization is distinct from the inflammatory infiltrate seen in acute or chronic inflammation, which would typically present with a dense lymphocytic or neutrophilic presence. The findings are less consistent with fibrosis, as there is no significant collagen deposition or smooth muscle thickening visible in the stroma. Therefore, the observed nuclear atypia and stratification within the epithelial layer strongly support the diagnosis of epithelial hyperplasia with keratinization.</think><answer>B</answer>

Figure 14: Failure case on histopathology recognition.

**You are an expert medical AI constructing a visual question answering dataset.**
**Generate exactly ONE question-answer pair.**

**Ground truth:**
Category: "{category}"
Description: "{description}"
Target bounding box: {box}

**Required task:**
Dimension: {dimension}
Question type: {question type}

**Critical rules:**
- The question must be answerable only from the medical image.
- Do not use patient history, symptoms, demographic information, laboratory results, or any non-visual clinical cue.
- Do not invent visual attributes that are not provided by the annotation.
- Keep the target entity semantically identical to the ground-truth category and description.

**Task-specific rules:**
- If Dimension = Anatomy Identification, include the exact text "{box}" in the question so that the model must inspect the specified region.
- If Dimension = Disease_and_Abnormality, do not include coordinates in the question; ask about the image-visible pathological finding.
- If Dimension = Visual Grounding, do not include coordinates in the question; ask the model to locate the target described by the ground truth.
- If Dimension = Modality Recognition, ask what imaging modality or technique was used, without revealing the correct modality in the question.

**Answer and option rules:**
- For open-ended questions, set A, B, C, and D to "None".
- For open-ended anatomy or abnormality questions, the answer must be only the core anatomy or pathology name, without modality words.
- For open-ended modality questions, the answer must be only the modality name.
- For open-ended visual grounding questions, the answer must be exactly {box}.
- For closed-ended visual grounding questions, use exactly the following options:
A: {option A}
B: {option B}
C: {option C}
D: {option D}
Answer: {correct letter}
- For other closed-ended questions, provide four medically plausible but mutually exclusive options and answer with one letter from A to D.

**Return a JSON object with exactly these fields:**
{
"question type": "open ended" | "closed ended", "question dimension": "{dimension}", "question": "...",
"A": "...",
"B": "...",
"C": "...",
"D": "...",
"answer": "..."
}

Figure 15: Prompt for VQA Construction.

**You are an expert medical AI quality inspector evaluating a generated visual question answering pair.**

**Absolute facts:**
The ground-truth category, description, and target region are accurate and strictly mapped to the provided image. Do not diagnose the image yourself and do not question whether the ground truth matches the image. Your only job is to evaluate whether the generated QA pair correctly and safely uses the ground truth without adding unsupported content.

Task dimension: {dimension}

Ground truth:
Category: {category}
Description: {description}

Generated QA pair:
Question: {question}
Options:
{options}
Answer: {gold answer}

**Evaluation criteria:**
1. QA-GT fidelity: the question and answer must faithfully translate the ground truth without semantic conflict.
2. No visual hallucination: the question must not introduce unsupported visual features or spatial relations.
3. Entity presence: for visual grounding, the question must clearly ask for the core anatomical structure or disease in the ground truth. For other tasks, output true.
4. No information leakage: for identification and modality tasks, the question must not reveal the answer. For other tasks, output true.
5. Distractor validity: for multiple-choice questions, distractors must be medically plausible but incorrect for this image. For open-ended questions, output true.

**Evaluate concisely. If all criteria are satisfied, output KEEP. If any criterion fails, output REJECT. Return JSON only:**
{
"reasoning": "...",
"image ground truth alignment": true | false,
"no visual hallucination": true | false,
"entity presence": true | false,
"no leakage": true | false,
"distractors valid": true | false,
"final decision": "KEEP" | "REJECT"
}

Figure 16: Prompt for VQA Verification.

**Reasoning style: short**
Vision mode: {full | image only}
Question type: {question type}
Modality: {modality}
Sequence: {sequence}
Dataset source: {dataset source}
Question: {question}
Gold answer: {gold answer}
Options:
{options}
Allowed modality-specific evidence terms: {modality lexicon}

**Internal schema to cover inside <think>:**
1. Localization: align the question focus with the visible ROI.
2. Appearance: describe one or two modality-specific findings.
3. Cross-view support: mention mask or DINO only when it adds concrete support.
4. Clinical decision: connect the visible evidence to the answer.

**Hard constraints:**
- Write 90-150 words inside <think>, targeting about 120 words.
- Use 3-4 sentences.
- Do not narrate your reasoning process.
- Use only image-visible evidence.
- Do not hallucinate patient history, lab values, or unseen slices.
- Do not repeat any sentence or restate the same evidence in new wording.
- Keep the answer tag exactly equal to the provided answer.
- Avoid generic phrases such as "combined visual cues", "semantic boundary", or "structural anomaly".
- Output one continuous rationale paragraph inside <think>.

**Output format:**
<think>medically grounded rationale</think><answer>{gold answer}</answer>

Figure 17: Prompt for Short-Style Reasoning Trace Generation.

**Reasoning style: long**
Vision mode: {full | image only}
Question type: {question type}
Modality: {modality}
Sequence: {sequence}
Dataset source: {dataset source}
Question: {question}
Gold answer: {gold answer}
Options:
{options}
Allowed modality-specific evidence terms: {modality lexicon}

**Internal schema to cover inside <think>:**
1. Localization: identify the ROI or structure relevant to the question.
2. Appearance: describe modality-specific findings visible in the original image.
3. Cross-view support: mention mask or DINO only when it adds concrete support.
4. Clinical decision: connect the visible evidence to the answer and briefly reject the strongest distractor or nearest alternative diagnosis.

**Hard constraints:**
- Write 240-380 words inside <think>, targeting about 300 words.
- Use 5-6 sentences.
- Include one explicit comparative phrase such as "favors ... over ...", "rather than ...", or "less consistent with ...".
- Do not narrate your reasoning process.
- Do not repeat any sentence or restate the same evidence in new wording.
- Keep the answer tag exactly equal to the provided answer.
- Do not hallucinate patient history, lab values, or unseen slices.
- Avoid generic filler and unsupported cross-view claims.
- Output one continuous rationale paragraph inside <think>.

**Output format:**
<think>medically grounded rationale</think><answer>{gold answer}</answer>

Figure 18: Prompt for Long-Style Reasoning Trace Generation.

**System:**
You are a medical judge-and-rewrite module. Evaluate whether a candidate rationale is medically grounded, modality-specific, structurally correct, and sufficiently differential when required. Do not explain your reasoning. Return one compact JSON object only:
{"score": float, "pass": bool, "rewrite needed": bool, "flags": [str], "summary": str}

**User:**
Style: {style}
Vision mode: {full | image only}
Question type: {question type}
Modality: {modality}
Question: {question}
Gold answer: {gold answer}
Needs differential reasoning: {true | false}
Rule-based feedback: {rule feedback}

**Candidate rationale:**
{candidate text}

**Judge rubric:**
- correct tag format and exact answer tag
- word budget for the selected style
- modality-specific medical evidence
- concrete use of mask or DINO evidence if mentioned
- avoidance of generic filler language
- differential reasoning when required

Return JSON only.

Figure 19: Prompt for Reasoning Quality Judge.

**System:**
You are a medical judge-and-rewrite module. Write a fresh final rationale from scratch. Never mention the user, prompt, candidate rationale, judge, rewrite process, gold answer, or instructions. Output only the final tagged answer.

**User:**
Produce a clean final response from scratch using the images and task.

**Style: {style}**
Vision mode: {full | image only}
Question type: {question type}
Modality: {modality}
Sequence: {sequence}
Question: {question}
Gold answer: {gold answer}
Options:
{options}
Allowed modality-specific evidence terms: {modality lexicon}
Failure hints to fix: {failure hints}

**Hard constraints:**
- Do not mention the prompt, candidate, judge, rewrite, user, or gold answer.
- Do not copy or discuss the previous failed rationale.
- Follow the word and sentence budget for {style}.
- Include differential reasoning when required.
- Use only image-visible evidence; use mask or DINO only when it gives concrete support.
- Do not repeat any sentence or restate the same evidence twice.
- Output only the final tagged response.

Figure 20: Prompt for Reasoning Rewrite.

**System:**
You are a highly rigorous expert medical evaluator. Your task is to assess the accuracy of an AI model's prediction against a verified ground truth.

**User:**
Compare the generated Prediction with the Ground Truth for the given Question.

[Question]: {question}
[Ground Truth]: {ground truth}
[Prediction]: {prediction}

**Evaluation criteria:**
1. Factual equivalence: the prediction must convey the same core medical meaning as the ground truth.
2. Synonyms and abbreviations: valid medical synonyms, acronyms, and standard alternative phrasings are acceptable.
3. Verbosity and contradictions: extra information is acceptable only if it is medically correct and does not contradict the ground truth.
4. Missing information: if the prediction misses critical clinical components required by the ground truth, it is incorrect.

**Output format:**
<think>
Step-by-step reasoning analyzing the match.
</think>
<judge>
1 or 0
</judge>

Figure 21: Prompt for Open-Ended Answer Equivalence Judge.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the main contributions, including the formulation of visual de-anchoring and spatial-semantic binding breakdown, the MedVIGOR framework, MedVIGOR-Bench, and empirical validation. The claims are supported by the method description and experiments, and the scope is further qualified in the limitations section.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the limitations in Section G.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: This paper is primarily based on empirical method design, benchmark construction, and experimental validation, and does not present theoretical results, formal assumptions, or mathematical proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper describes the model backbone, training stages, data construction pipeline, evaluation protocol, baselines, metrics, and additional implementation details in the appendix. These details provide a basis for reproducing or verifying the main experimental results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The data and code will be publicly available.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the base model, training stages, sample counts, learning rates, epochs, hardware, baselines, prompts, scoring protocol, and task-level benchmark composition. Additional training configurations are provided in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper reports deterministic accuracy-based comparisons but does not include error bars, confidence intervals, or statistical significance tests. Repeating all training and evaluation runs is computationally expensive due to multi-stage medical VLM training and large-scale multi-model evaluation, and we therefore focus on reporting detailed evaluation protocols and reproducibility settings.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Sufficient information on the compute resources is provided in Appendix D

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The authors have reviewed the NeurIPS Code of Ethics and believe the research conforms to it.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: MedVIGOR has the potential to support more reliable and interpretable medical VLM research by improving the grounding of model answers in case-specific visual evidence. MedVIGOR-Bench can also help evaluate whether existing medical VLMs rely on visual evidence or language priors, providing a useful resource for future model development. At the same time, we discuss potential risks in medical deployment, including over-reliance on model outputs and the need for expert validation before any real-world clinical use.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: Our paper poses no such risks.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite the existing datasets, models, and tools used to build MedVIGOR.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The paper introduces MedVIGOR-Bench and the MedVIGOR model. The benchmark composition, task definitions, evaluation protocol, prompts, and limitations are documented in the paper and appendix.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We use LLMs as key components for reasoning-supervision construction and open-ended evaluation, including trace generation, judging, rewriting, and answer-equivalence assessment. The corresponding usage and prompts are described in the method, evaluation, and prompt sections.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.