QGR: QUERY GROUNDING AND REFINEMENT FOR MEDICAL VISUAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual grounding links textual expressions to relevant image regions, while grounded reasoning requires model predictions and explanations to be supported by visual evidence. This remains challenging for medical vision-language models (VLMs), which can produce plausible explanations that are poorly aligned with disease relevant regions, leading to hallucinated reasoning and unreliable interpretations. Existing medical visual question answering (VQA) approaches mainly rely on supervised fine-tuning (SFT) or reinforcement learning (RL) to improve answer generation and reasoning, but provide limited direct supervision for question aware spatial grounding. As a result, spatial representations are often learned indirectly through the language-modeling objective, making it difficult to align reasoning with relevant visual evidence. We propose QGR (Query Grounding and Refinement), a multimodal grounding framework that connects language reasoning with object level visual evidence. QGR has two key components: Question Conditioned Query Refinement, which adapts object queries to the question context to capture relevant evidence, and Query Level Grounding Supervision, which directly guides their semantic assignment and localization quality. Together, they establish question aware semantic spatial correspondence at the query level. QGR jointly learns answer generation, reasoning, semantic region prediction, bounding-box localization, and query level grounding. Experiments show consistent improvements over the corresponding SFT baselines. On Med4VQA, QGR improves Closed accuracy by +3.20–17.64 points, Box localization by +7.64–52.15 points, and Final scores by +4.18–10.11 points across four imaging modalities. The gains also transfer out of domain, reaching +14.49 Final points on SLAKE MRI and up to +35.00 Box points on GEMeX. These results highlight the benefit of query level supervision for spatially grounded medical reasoning. Full training details and anonymized code and dataset are provided for reproducibility.

1 INTRODUCTION

Reliable interpretation of medical images requires identifying clinically relevant visual findings and interpreting them within their anatomical context. A medical vision language model (VLM), however, can generate a clinically correct response yet ground it on irrelevant image regions. A coherent clinical explanation does not, by itself, establish that the model has identified the appropriate visual evidence; similarly, accurate localization alone does not establish that the localized finding has been interpreted correctly Tascon-Morales et al. (2023); Nguyen et al. (2025b). Reliable medical VQA therefore requires *joint reasoning and grounding*, in which clinical interpretation guides the selection of visual evidence and the selected evidence informs the explanation Yao (2025); Zechen et al. (2024); Nguyen et al. (2025b). For questions concerning localized abnormalities, this requires consistency among the clinical answer, the findings used to justify it, and the spatial locations reported in the response.

Existing research provides several foundations for connecting these capabilities. Language-conditioned detectors such as MDETR and Grounding DINO associate textual queries with image regions Kamath et al. (2021); Liu et al. (2024), while VLM-R1 applies reinforcement learning to referring expression comprehension and object detection Shen et al. (2025). More recent approaches explicitly incorporate regional evidence into reasoning: VLM-R^3 interleaves textual reasoning with the selection and inspection of image regions Jiang et al. (2026), and the medical Localize-before-

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

Answer framework uses localized pathological regions to guide subsequent answer generation Nguyen et al. (2025b). Within medical reasoning models, QoQ-Med Dai et al. (2026) generates clinical reasoning alongside bounding-box annotations and diagnostic predictions, while GEMeX-RMCoT Liu et al. (2025) introduces intermediate reasoning steps grounded in relevant image regions Liu et al. (2025). Complementary developments in medical phrase grounding and the GEMeX Liu et al. (2025) benchmark provide methods and evaluation resources for associating clinical statements with spatial evidence Zhang et al. (2026); Liu et al. (2025).

Achieving this consistency calls for representations that capture both clinical meaning and spatial detail. General-purpose VLMs learn multimodal representations that support semantic interpretation, instruction following, and language generation, while dedicated detection architectures explicitly learn object-level representations through detection and bounding-box objectives Carion et al. (2020); Zhu et al. (2020); Liu et al. (2024). These complementary capabilities are particularly relevant to medical imaging, where pathological findings may be small, subtle, or difficult to distinguish from surrounding anatomy Tascon-Morales et al. (2023); Nguyen et al. (2025b). Grounded VLMs already support region and coordinate generation Bai et al. (2023); Peng et al. (2023); Chen et al. (2023), and recent approaches incorporate spatial evidence through region-aware reasoning and reinforcement learning Shen et al. (2025); Dai et al. (2026); Liu et al. (2025). Building on these developments, we investigate how the semantic representations of a VLM can interact directly with the object-level representations of a specialized detector. The aim is to make this interaction sensitive to the clinical question, so that semantic context guides the selection of relevant spatial evidence and the resulting localized representations inform the model's reasoning and response.

To this end, we propose **QGR (Query Grounding and Refinement)**, a unified framework that brings together the reasoning capabilities of a large VLM with the object level spatial representations of Deformable DETR Zhu et al. (2020). QGR is built around two main components: **Question Conditioned Query Refinement** and **Query Level Grounding Supervision**. In Question-Conditioned Query Refinement, multi-head cross-attention incorporates the question context into image-derived object queries, helping the model identify visual evidence that is relevant to the question. Query Level Grounding Supervision then provides direct supervision for these refined queries through semantic query assignment and localization quality objectives, encouraging them to represent the correct semantic regions and accurately localize the corresponding evidence. In this way, QGR moves beyond simply combining VLM and detector features: it explicitly connects *what the question asks* with *where the supporting evidence appears in the image*, establishing semantic–spatial correspondence at the query level. The overall QGR architecture and both components are described in detail in Section 3.

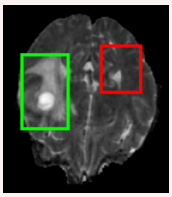
Contributions. Our main contributions are as follows:

- We propose **QGR (Query Grounding and Refinement)**, a unified framework that integrates multimodal semantic representations from a large VLM with image conditioned object query features from a specialized detector for grounded medical VQA.
- We formulate **medical VQA as a query-grounded reasoning problem**, where object query features are explicitly conditioned on question semantics and grounded through semantic query assignment and localization quality supervision. This establishes question conditioned semantic spatial correspondence at the query level, beyond conventional output level grounding.
- Extensive experiments on the in-domain Med4VQA benchmark and out-of-domain SLAKE and GEMeX benchmarks evaluate QGR across answer prediction, reasoning, semantic-region grounding, spatial localization, and zero-shot transfer.

2 Related Work

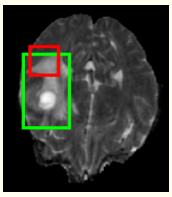
Vision-Language Models and Visual Reasoning. Recent multimodal large language models (MLLMs) have advanced visual understanding and question answering by jointly modeling visual and textual inputs. Models such as InternVL Chen et al. (2024), Eagle Shi et al. (2025), and Cambrian-1 Tong et al. (2024) strengthen visual perception through scalable vision encoders, mixtures of complementary visual encoders, and vision centric model design, respectively. While recent VLMs increasingly incorporate reasoning oriented training and post-training, general purpose MLLMs are

MedGemma



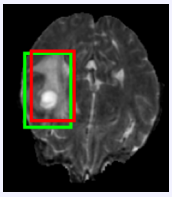
Question: Is the edema hyperintense or hypointense?
<answer>The edema is hyperintense.</answer>
<reason>Diagnosis: Edema. Key radiographic findings indicate edema present with localization at [134,56,170,98]. Clinical interpretation shows findings consistent with edema and localized to the annotated region.</reason>
<visual_regions>["edema"]</visual_regions>
<visual_loca>[[134,56,170,98]]</visual_loca>
Box Score: 0.00

Qwen3



Question: Is the edema hyperintense or hypointense?
<answer>The edema is hyperintense.</answer>
<reason>Diagnosis: Edema. Key radiographic findings indicate edema present with localization at [68,57,94,83]. Clinical interpretation shows findings consistent with edema and localized to the annotated region.</reason>
<visual_regions>["edema"]</visual_regions>
<visual_loca>[[68,57,94,83]]</visual_loca>
Box Score: 0.16

QGR (Ours)



Question: Is the edema hyperintense or hypointense?
<answer>The edema is hyperintense.</answer>
<reason>Diagnosis: Edema. Key radiographic findings indicate edema present with localization at [62,69,111,134]. Clinical interpretation shows findings consistent with edema and accurately aligned with the ground-truth region.</reason>
<visual_regions>["edema"]</visual_regions>
<visual_loca>[[68,62,107,125]]</visual_loca>
Box Score: 0.90

GPT-4 as Judge: GPT-4 evaluates open-ended answer and reasoning quality, while spatial localization is evaluated separately using bounding-box IoU.

Figure 1: **Qualitative comparison for edema localization in MRI.** Green boxes denote ground-truth regions and red boxes denote predicted regions. Model outputs follow the structured <answer>, <reason>, <visual_regions>, and <visual_loca> format. Although all models predict the correct answer, QGR achieves the strongest spatial alignment with the ground-truth edema region and the highest Box Score.

not specifically designed to provide explicit query-level supervision for anatomical region assignment and localization quality in grounded medical reasoning.

Visual Grounding and Medical VQA. Visual grounding connects linguistic expressions with corresponding visual regions and has expanded toward generalized grounding and open vocabulary detection Xiao et al. (2025). Methods such as MDETR Kamath et al. (2021), Grounding DINO Liu et al. (2024), Shikra Chen et al. (2023), and RegionGPT Guo et al. (2024) support language-conditioned detection, referring grounding, and region-level multimodal understanding. In the medical domain, VinDr-CXR-VQA Nguyen et al. (2025a) and GEMeX Liu et al. (2025) provide medical VQA data with spatially grounded visual evidence, while uMedGround Zou et al. (2025) addresses medical diagnostic phrase grounding by identifying diagnostic phrases and localizing their corresponding image regions. These works highlight the importance of spatial evidence in medical vision-language tasks. In contrast to output-level grounding or phrase region alignment, our approach explicitly uses question conditioned detector queries as grounding representations and supervises them through semantic query assignment and localization quality objectives.

Grounded Reasoning and RL Alignment. Chain-of-Thought (CoT) prompting Wei et al. (2022) encourages explicit intermediate reasoning in language models. Multimodal-CoT Zhang et al. (2023) incorporates visual and textual information into intermediate rationale generation, while Visual-CoT Rose et al. (2023) uses vision-language grounding to introduce visual information into multi-step reasoning. More recent grounded reasoning methods such as ARGUS Man et al. (2025) use object-

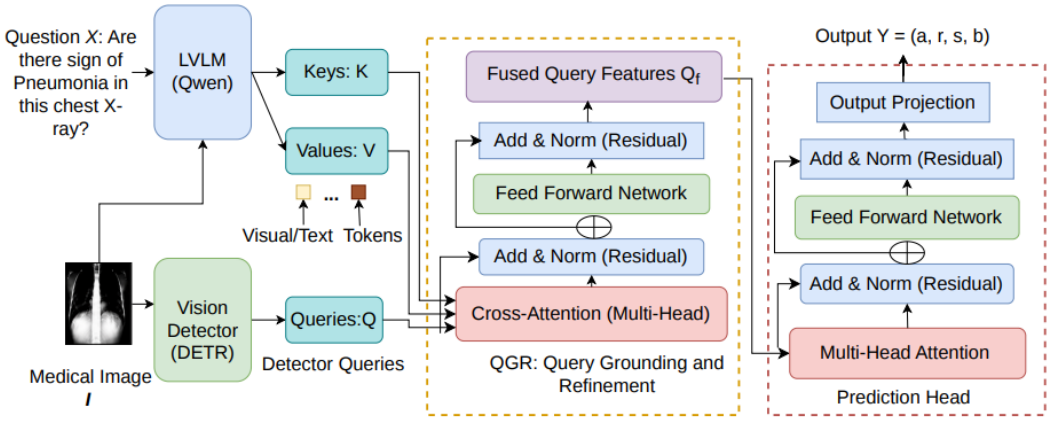


Figure 2: Overview of the proposed Query Grounding and Refinement (QGR) framework.

centric grounding as visual chain-of-thought signals for vision-centric multimodal reasoning. In the medical domain, Med-R1 Lai et al. (2026) and QoQ-Med Dai et al. (2026) investigate reinforcement learning-based post-training for medical vision-language reasoning. Preference optimization methods such as RRPO Sarkar & Etemad (2026) further show that large video language models can be self-aligned using preferred and non preferred responses constructed around common vision-language and spatio-temporal failure patterns, although RRPO targets video language understanding rather than grounded medical VQA.

3 Methodology

3.1 Preliminaries

Given a medical image I and a question $X = \{x_1, \ldots, x_T\}$ with T tokens, QGR predicts

$$Y = (a, r, s, b),$$

where a, r, s, and b denote the answer, reasoning, semantic region, and bounding box localization, respectively. QGR extracts multimodal semantic representations from (I, X) using a VLM and object level spatial representations from I using Deformable DETR. These representations are projected into a shared feature space and jointly decoded by a shared task module. In parallel, detector queries are conditioned on the question and supervised for semantic assignment and localization quality, enabling question-conditioned semantic spatial grounding.

As illustrated in Fig. 2, the VLM produces multimodal hidden features

$$H_v \in \mathbb{R}^{B \times N_v \times d_v},$$

where B, N_v, and d_v denote the batch size, number of multimodal tokens, and VLM hidden dimension, respectively. The detector processes only I and produces image conditioned object query features

$$H_o \in \mathbb{R}^{B \times N_o \times d_o},$$

where N_o and d_o denote the number of object queries and detector hidden dimension, respectively. Each object query j is associated with a normalized detector box $\mathbf{b}_j = (x_j, y_j, w_j, h_j)$, where (x_j, y_j) denotes the box center and (w_j, h_j) its width and height. Thus, H_v captures multimodal semantic information, whereas H_o provides object-level spatial evidence used for grounding and localization.

3.2 Query Grounding and Refinement

Question Conditioned Query Refinement. The detector identifies candidate object regions independently of the question. QGR conditions these object-query features on multimodal context before grounding. We first project the VLM and detector features to a common hidden dimension d:

$$F_v = P_v(H_v), \qquad Q_o = P_o(H_o). \tag{1}$$

Here, $F_v \in \mathbb{R}^{B\times N_v\times d}$ and $Q_o \in \mathbb{R}^{B\times N_o\times d}$ are the projected VLM and object-query features, respectively. They are concatenated for shared decoding:

$$F = \mathrm{Concat}(F_v, Q_o) \in \mathbb{R}^{B\times(N_v+N_o)\times d}. \tag{2}$$

The shared decoder operates on F for answer, reasoning, semantic region, and box-presence prediction.

For query refinement, the multimodal context $h_q = F_v[:, 0, :] \in \mathbb{R}^{B\times d}$ conditions each object query through cross-attention. For attention head h:

$$\mathcal{Q}^{(h)} = Q_o W_Q^{(h)}, \quad \mathcal{K}^{(h)} = h_q W_K^{(h)}, \quad \mathcal{V}^{(h)} = h_q W_V^{(h)}. \tag{3}$$

The corresponding cross-attention is

$$A^{(h)} = \mathrm{softmax}\left(\frac{\mathcal{Q}^{(h)}\mathcal{K}^{(h)\top}}{\sqrt{d_h}}\right)\mathcal{V}^{(h)}, \tag{4}$$

where d_h is the attention head dimension. The head outputs are concatenated and projected to obtain the question conditioned attention output A.

Finally, residual connections, layer normalization, and a feed-forward network refine the object query representation:

$$U = \mathrm{LN}(Q_o + A), \qquad Q_f = \mathrm{LN}(U + \mathrm{FFN}(U)). \tag{5}$$

Here, $Q_f \in \mathbb{R}^{B\times N_o\times d}$ denotes the refined object query features. As defined in Eqs. 3–5, each query incorporates multimodal context while retaining its associated detector box. Thus, QGR refines the semantic representation of each object query without directly modifying its predicted box coordinates.

Query-Level Grounding Supervision. Question-conditioned refinement alone does not guarantee that a refined query corresponds to the relevant semantic region or is spatially reliable. We, therefore, directly supervise each $Q_{f,j}$ through an assignment head and a quality head:

$$s_j = \mathrm{AssignHead}(Q_{f,j}), \qquad v_j = \mathrm{QualityHead}(Q_{f,j}), \tag{6}$$

where s_j contains logits over the semantic regions and a background class, and v_j is a scalar localization-quality logit.

Let $\{(b_n^{gt}, c_n)\}_{n=1}^{N_{gt}}$ denote the ground-truth boxes and their semantic-region labels. We match each detector query to the ground-truth box with the largest intersection over union (IoU):

$$u_j = \max_{1\le n\le N_{gt}} \mathrm{IoU}(b_j, b_n^{gt}), \tag{7}$$

Let c_j^* denote the semantic-region label of the ground-truth box achieving the maximum IoU for query j. This matching is performed independently for each query; several queries may receive the same region label. The assignment and quality targets are:

$$y_j^{\mathrm{asg}} = \begin{cases} c_j^*, & u_j > 0.1 \\ R, & \text{otherwise,} \end{cases} \qquad y_j^{\mathrm{qual}} = \begin{cases} 1, & u_j \ge 0.5 \\ 0, & \text{otherwise,} \end{cases} \tag{8}$$

where R denotes the background class. For examples with no ground-truth boxes, we set $u_j = 0$ and assign all queries to background. The lower assignment threshold allows partially overlapping queries to receive the semantic label of their matched region, whereas the stricter quality threshold identifies queries with stronger spatial alignment.

The corresponding query level grounding losses are

$$\mathcal{L}_{\mathrm{assign}} = \mathrm{CE}(s, y^{\mathrm{asg}}), \qquad \mathcal{L}_{\mathrm{quality}} = \mathrm{BCEWithLogits}(v, y^{\mathrm{qual}}), \tag{9}$$

where CE and $\mathrm{BCEWithLogits}$ denote cross-entropy and binary cross-entropy with logits, respectively. These losses encourage the refined queries to encode both the semantic identity and spatial reliability of question relevant visual evidence. The complete QGR procedure is summarized in Algorithm 1.

Training Objective. QGR-1H (final model) uses a single shared decoder, its shared representation supports four prediction functions: answer, reasoning, semantic region, and box localization. These

Algorithm 1: Query Grounding and Refinement (QGR)

Input: VLM features F_v, object query features Q_o, predicted boxes $\{b_j\}_{j=1}^{N_o}$, GT regions $\{(b_n^{gt}, c_n)\}_{n=1}^{N_{gt}}$
Output: Refined queries Q_f, $\mathcal{L}_{\text{assign}}$, $\mathcal{L}_{\text{quality}}$
$h_q \leftarrow F_v[:, 0:1, :]$
$A \leftarrow \text{MHA}(Q_o, h_q, h_q)$
$U \leftarrow \text{LN}(Q_o + A)$
$Q_f \leftarrow \text{LN}(U + \text{FFN}(U))$
$(s, v) \leftarrow (\text{AssignHead}(Q_f), \text{QualityHead}(Q_f))$
for $j = 1, \ldots, N_o$ **do**
  $u_j \leftarrow \max_n \text{IoU}(b_j, b_n^{gt})$
  $c_j^* \leftarrow$ class of the best-matched GT box
  $y_j^{\text{asg}} \leftarrow c_j^*$ if $u_j > 0.1$, else R
  $y_j^{\text{qual}} \leftarrow 1$ if $u_j \geq 0.5$, else 0
$\mathcal{L}_{\text{assign}} \leftarrow \text{CE}(s, y^{\text{asg}})$
$\mathcal{L}_{\text{quality}} \leftarrow \text{BCEWithLogits}(v, y^{\text{qual}})$
return $Q_f, \mathcal{L}_{\text{assign}}, \mathcal{L}_{\text{quality}}$

predictions are jointly optimized with the detector and query-level grounding objectives. We group the training losses according to their role as

$$\mathcal{L}_{\text{total}} = \underbrace{\mathcal{L}_{\text{LM}}}_{\text{Qwen}} + \underbrace{\mathcal{L}_{\text{Det}}}_{\text{DETR}} + \underbrace{\mathcal{L}_{\text{Reg}} + \mathcal{L}_{\text{Box}}}_{\text{Shared decoder auxiliaries}} + \underbrace{\mathcal{L}_{\text{Asg}} + \mathcal{L}_{\text{Qual}}}_{\text{QGR query supervision}} . \tag{10}$$

Here, $\mathcal{L}_{\text{LM}}$ supervises answer and reasoning generation, $\mathcal{L}_{\text{Det}}$ trains the detector, $\mathcal{L}_{\text{Reg}}$ and $\mathcal{L}_{\text{Box}}$ supervise region and box prediction from the shared decoder representation, and $\mathcal{L}_{\text{Asg}}$ and $\mathcal{L}_{\text{Qual}}$ supervise query assignment and localization quality, respectively.

4 Experiments

In this section we summarize the training strategy, baseline models, evaluation datasets, and evaluation metrics used in our experiments. Our experiments aim to answer the following questions: (i) Does explicit query-level grounding improve both semantic reasoning and spatial grounding in medical VQA? (ii) How does the proposed QGR framework compare with strong zero-shot VLMs, medical-domain VLMs, and Grounding SFT baselines? (iii) How well does QGR generalize across in-domain and out-of-domain datasets and imaging modalities?

4.1 Experimental Setup

Training. We train five model configurations using Grounding Supervised Fine-Tuning (SFT): VLM-R1-OVD, Qwen3-VL-4B, MedGemma, and QGR with either Qwen3-VL-4B or MedGemma as the backbone. This provides backbone matched comparisons between QGR and the corresponding SFT models, with VLM-R1-OVD included as an additional grounding baseline. GRPO is evaluated separately as a post-training ablation.

Grounding SFT. QGR jointly optimizes the VLM and Deformable DETR branch under supervision for answer generation, reasoning, semantic-region prediction, and bounding-box localization. In addition, semantic query assignment and query quality objectives directly supervise the question conditioned object queries, encouraging semantic relevance and accurate localization of the supporting visual evidence. QGR training follows Algorithm 1, with implementation details provided in Appendix B.

Baseline Models. We compare QGR with three categories of baselines: (1) zero-shot foundation VLMs, including BLIP2-OPT-2.7B Li et al. (2023b), LLaVA-1.5-7B Liu et al. (2023), and Qwen3-VL-4B Yang et al. (2025); (2) medical domain VLMs, including MedGemma Sellergren et al. (2025), QoQ-Med-VL-7B Dai et al. (2026), and LLaVA-Med Li et al. (2023a); and (3) SFT trained baselines, including Qwen3-VL-4B, MedGemma, and VLM-R1-OVD Shen et al. (2025). Together, these comparisons evaluate reasoning and spatial grounding across general purpose, medical domain, and grounding oriented models. GRPO results are reported separately in the ablation study.

Table 1: Comparison of QGR with zero-shot, SFT, and GRPO baselines on Med4VQA (%). We report modality wise closed question accuracy (Closed), bounding-box localization (Box), and Final scores. QGR is evaluated with both Qwen3-VL-4B Yang et al. (2025) and MedGemma Sellergren et al. (2025) backbones.

Model	CT			MRI			Ultrasound			X-ray		
	Closed	Box	Final	Closed	Box	Final	Closed	Box	Final	Closed	Box	Final
Other Baselines												
BLIP2-OPT-2.7B (Zero-shot)	8.40	0.00	5.87	12.00	0.00	6.40	2.19	0.00	4.52	0.00	0.00	4.02
LLaVA-1.5-7B (Zero-shot)	9.24	0.00	8.90	12.00	0.00	9.33	7.30	0.00	7.08	7.87	0.00	8.23
LLaVA-Med (Zero-shot)	3.36	0.00	4.72	20.00	0.00	8.00	7.30	0.00	5.48	10.24	0.00	6.07
QoQ-Med-VL-7B (Zero-shot)	13.45	7.38	28.45	21.60	11.47	29.53	8.76	12.33	26.23	2.36	6.02	24.13
VLM-R1-OVD (Zero-shot)	7.56	6.24	23.20	36.00	7.78	26.58	20.44	7.26	23.75	19.69	3.88	21.61
+ SFT	6.72	7.00	25.97	34.40	10.01	38.05	21.17	15.68	29.14	26.77	5.94	30.52
+ GRPO	6.72	6.80	25.92	35.20	10.14	38.30	21.17	15.07	28.59	26.77	5.75	30.58
MedGemma (Zero-shot)	10.08	0.73	14.13	8.80	0.87	9.73	4.38	1.19	11.81	2.36	1.49	13.87
+ SFT	**78.15**	2.45	47.92	92.80	4.72	70.96	**84.67**	40.50	75.34	89.76	8.73	70.72
+ GRPO	77.31	2.37	47.58	89.60	5.04	69.87	**84.67**	40.08	74.81	90.55	8.95	71.39
+ QGR	**78.15**	**1.60**	**49.77**	**93.60**	**32.80**	**75.92**	83.94	**97.20**	**86.23**	**92.13**	**34.40**	**77.27**
Qwen3-VL(Zero-shot)	14.29	0.88	24.73	19.20	0.05	26.01	8.03	2.15	29.39	7.09	7.71	26.74
+ SFT	69.75	8.76	52.11	90.40	18.12	73.82	70.80	47.05	77.75	78.74	23.79	75.11
+ GRPO	68.07	8.59	52.08	88.80	17.40	73.29	73.72	47.18	78.42	78.74	24.11	75.16
+ QGR	**87.39**	**16.40**	**56.29**	**93.60**	**56.40**	**80.40**	**84.67**	**99.20**	**87.25**	**94.49**	**70.00**	**85.22**

Datasets. We evaluate QGR on one in-domain benchmark and two out-of-domain benchmarks.

1. Med4VQA (In-domain). We curate the proposed Med4VQA dataset for training and in-domain evaluation. The raw medical images are collected from the VGMED dataset Liu et al. (2026) and are publicly available at VGMED Dataset. We further construct structured grounded VQA annotations comprising answer supervision, reasoning, semantic region labels, and bounding-box localization. Med4VQA contains 6,898 medical images and over 137K grounded VQA samples across four imaging modalities: CT, MRI, ultrasound, and X-ray. Each sample explicitly associates textual reasoning with localized visual evidence through structured annotations. Detailed dataset construction and statistics are provided in Appendix A.

2. SLAKE (Out-of-domain). To evaluate cross-dataset generalization, we directly evaluate models trained on Med4VQA on SLAKE Liu et al. (2021) without further finetuning. SLAKE is a medical VQA dataset containing radiology and clinical images with semantically labeled annotations, supporting both open ended and closed ended questions. It contains 642 images with 450 images (9,849 questions) for training, 96 images (2,109 questions) for validation, and 96 images (2,070 questions) for testing. We use the official test split for zero-shot evaluation. The dataset is publicly available at SLAKE Dataset. Its differences from Med4VQA in image distribution, question formulation, answer space, and annotation structure provide a complementary setting for evaluating cross dataset transfer.

3. GEMeX (Out-of-domain). We further evaluate cross dataset generalization on GEMeX-ThinkVG Liu et al. (2025). We construct a balanced evaluation subset of 80 VQA pairs from 11 unique chest X-ray images, comprising 20 closed ended, 20 open ended, 20 single-choice, and 20 multiple choice questions. All models are trained on Med4VQA and evaluated directly on this subset without GEMeX specific fine-tuning, providing an additional zero-shot evaluation of grounded reasoning and spatial localization under cross-dataset distribution shift.

Evaluation Metrics. We evaluate both semantic reasoning and spatial grounding. Closed-form questions use exact answer matching, while open-ended answers and reasoning are assessed using GPT-4o-mini as an LLM judge. Region prediction is measured using strict set-based F1, and localization is evaluated using IoU-based top-1 box matching. Final results are reported using Closed Accuracy, Open QA Score, Reason Score, Region Score, Box Score, and their average as the Final Score. Details of the metrics are provided in Appendix C

4.2 Results and Analysis

Tables 1 and 2 summarize the in-domain and out-of-domain results, respectively, with detailed results provided in Appendix C. As shown in Fig. 1, the qualitative comparison demonstrates QGR's superior spatial grounding, with more accurate alignment to the ground-truth edema region than the baseline

Table 2: Out-of-domain results on SLAKE and GEMeX (%). $^{\mathcal{I}}$ denotes models trained on Med4VQA and evaluated on the target dataset without additional target-domain training. We report bounding-box localization (Box) and Final scores on SLAKE, and open-question performance (Open), Box, and Final scores on GEMeX. QGR is evaluated with both Qwen3-VL-4B Yang et al. (2025) and MedGemma Sellergren et al. (2025) backbones.

	SLAKE						GEMeX		
	CT		MRI		X-ray		X-ray		
Model	**Box**	**Final**	**Box**	**Final**	**Box**	**Final**	**Open**	**Box**	**Final**
MedGemma	10.17	23.28	14.47	25.12	0.00	23.19	30.01	0.00	22.75
+ SFT$^{\mathcal{I}}$	10.17	**28.48**	14.47	**24.93**	0.00	26.29	34.32	2.50	16.56
+ QGR + SFT$^{\mathcal{I}}$	**10.59**	23.28	**30.70**	24.69	0.00	**29.14**	**34.36**	**37.50**	**21.27**
Qwen3-VL	10.17	38.58	14.47	45.88	0.00	31.75	29.60	2.50	34.13
+ SFT$^{\mathcal{I}}$	10.17	34.86	14.47	34.06	0.00	29.39	15.82	1.25	9.96
+ QGR + SFT$^{\mathcal{I}}$	**34.32**	**40.52**	**58.33**	**48.55**	**12.74**	**37.38**	**35.61**	**28.75**	**19.77**

models. QGR consistently improves over its corresponding SFT baseline on Med4VQA. With Qwen3-VL-4B, Closed accuracy increases by +3.20–17.64 points across the four modalities, while Box localization improves by +7.64–52.15 points. These gains lead to Final-score improvements of +4.18–10.11 points, with the strongest improvements in spatial grounding. In particular, QGR achieves Box scores of 99.20% on ultrasound and 70.00% on X-ray, highlighting the benefit of question conditioned query refinement and supervision.

This advantage also transfers out of domain. On SLAKE, QGR improves over Qwen3-VL-4B SFT by +5.66, +14.49, and +7.99 points in Final score on CT, MRI, and X-ray, respectively, while increasing Box scores from 10.17% to 34.32%, 14.47% to 58.33%, and 0.00% to 12.74%. On GEMeX, QGR improves Box localization from 1.25% to 28.75% with Qwen3-VL-4B and from 2.5% to 37.5% with MedGemma. The corresponding Final scores increase from 9.96% to 19.77% and from 16.56% to 21.27%, respectively.

Overall, the results consistently show that QGR's strongest advantage lies in spatial grounding. The gains across modalities, backbones, and out-of-domain datasets support the effectiveness of query level refinement and supervision for connecting the question with its supporting visual evidence. In contrast, GRPO provides only modest and modality dependent changes over QGR-SFT, indicating that the main gains come from the QGR grounding mechanism itself.

5 Ablation Study

We conduct an architecture-level ablation and a GRPO ablation on Med4VQA to disentangle the contributions of feature concatenation, Query Grounding and Refinement (QGR), decoder design, backbone choice, and GRPO training. As shown in Table 3 and Table 4, we compare six variants: (1) *Qwen3-VL*, the Grounding-SFT baseline that predicts structured answer, reasoning, region, and box outputs without an explicit detector; (2) *Qwen3-VL+DETR*, which introduces detector-based spatial features through feature concatenation without QGR; (3) *QGR-4H (Qwen3-VL)*, which combines feature concatenation and QGR with four task-specific decoder heads; (4) *QGR-1H (Qwen3-VL)*, which replaces the four task-specific decoder heads with a single shared decoder head; (5) *QGR-1H (MedGemma)*, which retains the same single-head QGR architecture while replacing the Qwen3-VL backbone with MedGemma; and (6) *QGR-GRPO*, which further optimizes QGR using Group Relative Policy Optimization (GRPO) with different training subset sizes.

The ablation reveals clear architectural trade-offs. Qwen3-VL exhibits strong language performance but relatively weak box localization, while incorporating DETR features substantially improves localization on several modalities at the cost of language-oriented metrics. Adding QGR provides a stronger balance between language reasoning and spatial grounding. Compared with QGR-4H, QGR-1H with Qwen3-VL improves Box localization on CT (16.40 vs. 9.20), ultrasound (99.20 vs. 98.80), and X-ray (70.00 vs. 60.00), while achieving higher Final scores on ultrasound (87.25 vs. 87.16) and X-ray (85.22 vs. 83.04). QGR-4H instead achieves higher Final scores on CT (56.53 vs. 56.29) and MRI (81.60 vs. 80.40), indicating modality-dependent trade-offs between shared and task-specific decoder heads. Finally, replacing Qwen3-VL with MedGemma in the QGR-1H architecture yields competitive Region grounding but lower Box and Final scores overall, showing that the gains of the proposed architecture are also influenced by the underlying VLM backbone.

Table 3: Architecture ablation on Med4VQA with Grounding SFT. *Feature Concat (FC)* denotes feature concatenation, *Query Grounding and Refinement (QGR)* denotes the proposed query-level grounding and refinement mechanism, and *Decoder Head (DH)* indicates the use of a decoder head.

Model	FC	QGR	DH	Modality	Closed	Open	Reason	Region	Box	Final
Qwen3-VL	✗	✗	✗	CT	69.75	61.94	50.42	66.80	8.76	51.53
				MRI	90.40	93.52	66.32	97.20	18.12	73.11
				Ultrasound	70.80	93.89	78.43	90.90	47.05	76.21
				X-ray	78.74	97.64	70.66	91.00	23.79	72.37
Qwen3-VL+DETR	✓	✗	✓	CT	48.99	37.91	34.04	20.80	15.60	31.47
				MRI	53.99	58.65	58.65	88.40	49.20	61.78
				Ultrasound	32.24	10.56	60.56	83.60	80.40	53.47
				X-ray	58.00	61.40	56.42	98.50	51.60	65.18
QGR-4H (Qwen3-VL)	✓	✓	✓	CT	87.39	64.76	54.89	66.40	9.20	56.53
				MRI	92.80	87.68	67.92	95.60	64.00	81.60
				Ultrasound	84.67	78.58	73.73	100.00	98.80	87.16
				X-ray	94.49	87.68	73.05	100.00	60.00	83.04
QGR-1H (Qwen3-VL)	✓	✓	✓	CT	87.39	65.00	55.20	57.47	16.40	56.29
				MRI	93.60	87.68	69.12	95.20	56.40	80.40
				Ultrasound	84.67	78.32	74.08	100.00	99.20	87.25
				X-ray	94.49	87.80	73.82	100.00	70.00	85.22
QGR-1H (MedGemma)	✓	✓	✓	CT	78.15	58.71	49.19	61.20	1.60	49.77
				MRI	93.60	88.80	67.60	96.80	32.80	75.92
				Ultrasound	83.94	77.08	72.92	100.00	97.20	86.23
				X-ray	92.13	87.15	72.67	100.00	34.40	77.27

Table 4: Comparison of QGR with the Qwen3-VL-4B backbone after GRPO training on 500, 300, and 200 sample subsets across Med4VQA modalities. Positive Δ indicates improvement from Subset300 to Subset500.

Metric	CT				MRI				Ultrasound				X-ray			
	500	300	200	Δ	500	300	200	Δ	500	300	200	Δ	500	300	200	Δ
Closed	86.55	85.71	**87.39**	+0.84	**95.20**	94.40	93.60	+0.80	**86.13**	84.67	84.67	+1.46	**94.49**	93.70	92.91	+0.79
Open	65.19	64.18	**65.65**	+1.01	**88.40**	87.68	86.40	+0.72	79.12	**81.68**	79.12	-2.56	87.32	86.99	**87.34**	+0.33
Reason	**54.66**	54.44	53.98	+0.22	**69.20**	68.80	68.48	+0.40	73.71	**74.93**	74.16	-1.22	**74.12**	73.25	72.59	+0.87
Region	57.33	**66.40**	66.00	-9.07	95.20	**95.60**	**95.60**	-0.40	**100.00**	**100.00**	**100.00**	+0.00	**100.00**	**100.00**	**100.00**	+0.00
Box	**15.20**	9.20	9.20	+6.00	57.20	**64.80**	63.60	-7.60	98.40	**98.80**	**98.80**	-0.40	**70.00**	60.80	59.60	+9.20
Final	55.79	55.99	**56.44**	-0.20	81.04	**82.26**	81.54	-1.22	87.47	**88.02**	87.35	-0.55	**85.18**	82.95	82.49	+2.23

The GRPO ablation in Table 4 shows more modest and modality-dependent effects. Compared with QGR-1H SFT, the 500-sample GRPO setting improves the Final score on MRI (80.40 to 81.04) and ultrasound (87.25 to 87.47), while slightly decreasing it on CT (56.29 to 55.79) and X-ray (85.22 to 85.18). Overall, these results suggest that GRPO does not uniformly improve over SFT, and its effectiveness depends on both the training subset size and imaging modality.

6 Conclusion and Limitations

We presented QGR (Query Grounding and Refinement) for grounded medical VQA, with the goal of making model reasoning more closely tied to the visual evidence in medical images. Rather than relying only on supervision at the final output, QGR brings question information directly into object level queries and provides explicit supervision for their semantic relevance and localization. Our experiments on Med4VQA, SLAKE, and GEMeX show that this design leads to consistent improvements in overall performance, with the largest gains observed in bounding-box localization. The results across different imaging modalities and datasets further suggest that the learned grounding is not limited to a single evaluation setting. QGR still has several limitations. Its training requires region level annotations, which can be costly to obtain for medical images, and its grounding performance remains dependent on the quality of the underlying detector. Errors in region annotations or weak detector predictions can therefore affect the evidence available to the model. Future work will investigate weaker forms of grounding supervision, more efficient integration of spatial information, and evaluation across broader clinical tasks and datasets.

REFERENCES

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*, 2(1):1, 2023.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

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

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

Qiushan Guo, Shalini De Mello, Hongxu Yin, Wonmin Byeon, Ka Chun Cheung, Yizhou Yu, Ping Luo, and Sifei Liu. Regiongpt: Towards region understanding vision language model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13796–13806. IEEE, 2024.

Chaoya Jiang, Yongrui Heng, Wei Ye, Haiyang Xu, Ming Yan, Ji Zhang, Fei Huang, and Shikun Zhang. Vlm-r^3: Region recognition, reasoning, and refinement for enhanced multimodal chain-of-thought. *Advances in Neural Information Processing Systems*, 38:63841–63869, 2026.

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

Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, Yuheng Li, Konstantinos Psounis, and Xiaofeng Yang. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *IEEE transactions on medical imaging*, 2026.

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, 2023a.

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

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)*, pp. 1650–1654. IEEE, 2021.

Bo Liu, Xiangyu Zhao, Along He, Yidi Chen, Huazhu Fu, and Xiao-Ming Wu. Gemex-rmcot: An enhanced med-vqa dataset for region-aware multimodal chain-of-thought reasoning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 13213–13220, 2025.

Guimeng Liu, Tianze Yu, Somayeh Ebrahimkhani, Shawn Lin, Kok Pin Ng, and Man Cheung. How do medical mllms fail? a study on visual grounding in medical images. In *International Conference on Learning Representations*, volume 2026, pp. 95215–95268, 2026.

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

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

Yunze Man, De-An Huang, Guilin Liu, Shiwei Sheng, Shilong Liu, Liang-Yan Gui, Jan Kautz, Yu-Xiong Wang, and Zhiding Yu. Argus: Vision-centric reasoning with grounded chain-of-thought. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14268–14280. IEEE, 2025.

Dang H Nguyen, Hieu H Pham, and Hao T Nguyen. Vindr-cxr-vqa: A visual question answering dataset for explainable chest x-ray analysis with multi-task learning. *arXiv preprint arXiv:2511.00504*, 2025a.

Dung Nguyen, Minh Khoi Ho, Huy Ta, Thanh Tam Nguyen, Qi Chen, Kumar Rav, Quy Duong Dang, Satwik Ramchandre, Son Lam Phung, Zhibin Liao, et al. Localizing before answering: A benchmark for grounded medical visual question answering. In *Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)*. International Joint Conferences on Artificial Intelligence Organization, 2025b.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

Daniel Rose, Vaishnavi Himakunthala, Andy Ouyang, Ryan He, Alex Mei, Yujie Lu, Michael Saxon, Chinmay Sonar, Diba Mirza, and William Yang Wang. Visual chain of thought: bridging logical gaps with multimodal infillings. *arXiv preprint arXiv:2305.02317*, 2023.

Pritam Sarkar and Ali Etemad. Self-alignment of large video language models with refined regularized preference optimization. *Advances in Neural Information Processing Systems*, 38:42633–42665, 2026.

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.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.

Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, Yilin Zhao, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, et al. Eagle: Exploring the design space for multimodal llms with mixture of encoders. In *International Conference on Learning Representations*, volume 2025, pp. 92628–92646, 2025.

Sergio Tascon-Morales, Pablo Márquez-Neila, and Raphael Sznitman. Localized questions in medical visual question answering. In *International conference on medical image computing and computer-assisted intervention*, pp. 361–370. Springer, 2023.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37: 87310–87356, 2024.

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

Linhui Xiao, Xiaoshan Yang, Xiangyuan Lan, Yaowei Wang, and Changsheng Xu. Towards visual grounding: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Yanyu Yao. A survey on multimodal large language models. In *Computer Science Undergradaute Conference 2025@ XJTU*, 2025.

Bai Zechen et al. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.

Wenjian Zhang, Shekhar S Chandra, and Aaron Nicolson. Generalised medical phrase grounding. *IEEE Transactions on Medical Imaging*, 2026.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

Ke Zou, Yang Bai, Bo Liu, Yidi Chen, Zhihao Chen, Yang Zhou, Xuedong Yuan, Meng Wang, Xiaojing Shen, Xiaochun Cao, et al. Uncertainty-aware medical diagnostic phrase identification and grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

APPENDIX

A MED4-GROUNDED VQA DATASET

We construct a grounded multimodal medical visual question answering dataset, termed **Med4-Grounded VQA**. The dataset is designed to evaluate both semantic medical reasoning and spatial grounding. Unlike standard medical VQA datasets that primarily focus on text answers, Med4-Grounded VQA explicitly links each answer to visual evidence through clinically relevant labels, segmentation masks, and bounding box annotations.

The dataset covers four major medical imaging modalities: CT, MRI, X-ray, and Ultrasound. In total, it contains 6,898 images. Each image is associated with clinically meaningful annotations and 10 question-answer pairs. During preprocessing, every question is expanded into both open-ended and closed-ended forms, resulting in 20 VQA samples per image. This produces 123,400 training samples and 14,560 test samples. The dataset includes 27 unique medical labels, and all VQA pairs are generated using GPT-based prompting to ensure question diversity and clinically meaningful reasoning.

A.1 DATASET OVERVIEW

Table 5 summarizes the train and test splits. The training split contains 6,170 images and 123,400 VQA samples, while the test split contains 728 images and 14,560 VQA samples. Each split is balanced between open-ended and closed-ended questions, ensuring that models are evaluated on both free-form reasoning and direct answer selection.

A.2 MODALITY DISTRIBUTION

The dataset is distributed across CT, MRI, X-ray, and Ultrasound images as shown in Table 6 . This enables evaluation across different image characteristics, anatomical structures, and disease appearances. CT and MRI contain diverse anatomical and pathological labels, while X-ray and Ultrasound focus on clinically important modality-specific findings.

A.3 LABEL NAMES

The dataset contains 27 unique medical labels across the four modalities as shown in Table 7. CT contains the largest set of labels, covering organs and anatomical structures such as airway, aorta, bladder, kidney, liver, lung, pancreas, spleen, and uterus. MRI includes labels such as edema, necrosis, lesion, prostate, and liver. X-ray focuses on pneumothorax, while Ultrasound includes breast lesion and thyroid nodule.

A.4 QUESTION TEMPLATES

For each modality, we design modality-specific question templates as shown in Table 8, 9, 10, and 11. These templates are intended to cover appearance, visibility, abnormality presence, size, localization, and disease identification. Each template is instantiated using the corresponding medical label. For closed-ended questions, explicit answer options are provided. For open-ended questions, the model must generate a natural-language response and provide clinically grounded reasoning.

A.5 STRUCTURED GROUNDING ANNOTATIONS

Each VQA sample is paired with structured grounding information. In addition to the answer and reasoning text, each sample includes a region label and a bounding box. This allows the dataset to support evaluation of both language-level correctness and image-level localization. In our experiments, model outputs are represented using the following structured format:

```
<answer>...</answer>
<reason>...</reason>
<visual_regions>["region_name"]</visual_regions>
```

Table 5: Overview of the Med4-Grounded VQA dataset.

Split	Images	VQA Count	Open	Closed	Per Image
Train	6170	123400	61700	61700	20
Test	728	14560	7280	7280	20

Table 6: Modality distribution in Med4-Grounded VQA.

Modality	Train Img	Test Img	Total Img	Train VQA	Test VQA	Total VQA
CT	1555	199	1754	31100	3980	35080
MRI	1599	158	1757	31980	3160	35140
X-ray	1499	189	1688	29980	3780	33760
Ultrasound	1517	182	1699	30340	3640	33980

`<visual_locations>[[x1,y1,x2,y2]]</visual_locations>`

This structured format allows deterministic evaluation of the region and box predictions while still supporting flexible semantic evaluation of open-ended answers and reasoning.

B Implementation and Training Details

Training environment. All experiments are conducted on a single GPU(L40S)using CUDA with PyTorch memory allocation configured by `PYTORCH_CUDA_ALLOC_CONF=expandable_segments:True`. We use one GPU for both Grounding SFT and GRPO training. Hugging Face model and dataset caches are stored locally, and Weights & Biases is used for experiment logging.

Grounding SFT. For the supervised stage, QGR is initialized with `Qwen/Qwen3-VL-4B-Instruct` as the language backbone and `SenseTime/deformable-detr` as the detector backbone. The model is trained on the Med4VQA training JSONL file with images loaded from the VGMED-derived Med4 image root. We train for one epoch with batch size 1 and gradient accumulation 8, giving an effective batch size of 8. The base learning rate is 5×10^{-6} with weight decay 0.01 and warmup ratio 0.05. The detector learning-rate multiplier is 3.0 and the task-head learning-rate multiplier is 5.0. The task hidden dimension is 256, with 4 task heads and dropout 0.1. LoRA is enabled with rank 16, alpha 32, and dropout 0.05. The supervised loss weights are $\lambda_{\text{text}} = 1.0$, $\lambda_{\text{det}} = 0.5$, region auxiliary weight 0.5, box auxiliary weight 0.25, query-assignment weight 0.35, and query-quality weight 0.20. Checkpoints are saved every 1000 steps with a maximum of two checkpoints retained.

Grounding-aware GRPO. The GRPO stage starts from the Grounding SFT checkpoint `checkpoint-15520`. We train on a subset of 200 samples for one epoch with batch size 1 and gradient accumulation 1. The learning rate is 1×10^{-7} with weight decay 0.01, warmup ratio 0.10, and maximum gradient norm 0.1. For each prompt, the policy samples 2 generations with maximum generation length 256, temperature 0.70, and top-p 0.90. The KL coefficient is $\beta_{\text{KL}} = 0.10$, and the auxiliary grounding loss weight is 1.5. The reference model is placed on CPU to reduce GPU memory usage.

Grounding and reward settings. During GRPO, detector filtering uses query threshold 0.03, NMS IoU threshold 0.30, maximum 5 detection boxes, minimum area ratio 5×10^{-5}, maximum area ratio 0.95, and box gate threshold 0.05. The reward combines answer, reasoning, format, region, box, detector-consistency, and detector-GT terms with weights 0.2, 0.1, 1.5, 1.5, 1.5, 0.5, and 0.5, respectively. Checkpoints are saved every 100 steps with a maximum of two checkpoints retained. We use random seed 123 for GRPO training.

Table 7: Medical label categories by modality.

Modality	Labels
CT	airway, aorta, bladder, bone, brain, colon, duodenum, esophagus, face, gallbladder, intestine, kidney, lesion, liver, lung, pancreas, sacrum, spleen, stomach, trachea, uterus
MRI	aorta, edema, esophagus, gallbladder, lesion, liver, necrosis, pancreas, prostate, spleen
X-ray	pneumothorax
Ultrasound	breast lesion, thyroid nodule

Table 8: CT question templates.

Open-ended	**Closed-ended**
1. Does the <LABEL> appear hypo- or hyper-dense?	1. Does the <LABEL> appear hypo- or hyper-dense? **Options:** hypodense, hyperdense, not specified
2. Is the <LABEL> homogeneous or heterogeneous?	2. Is the <LABEL> homogeneous or heterogeneous? **Options:** homogeneous, heterogeneous, not specified
3. Is the <LABEL> visible in the image?	3. Is the <LABEL> visible in the image? **Options:** yes, no
4. Is there a <LABEL> in the image?	4. Is there a <LABEL> in the image? **Options:** yes, no
5. Is there any additional abnormality besides the recorded finding?	5. Is there any additional abnormality besides the recorded finding? **Options:** yes, no
6. What abnormalities are present in this image?	6. What abnormalities are present in this image?
7. What is the size category of the dominant abnormality?	7. What is the size category of the dominant abnormality? **Options:** small, moderate, large, not specified
8. What is the size of the <LABEL>?	8. What is the size of the <LABEL>? **Options:** small, moderate, large, not specified
9. Where is the dominant abnormality located?	9. Where is the dominant abnormality located?
10. Which disease or <LABEL> is identified in this image?	10. Which disease or <LABEL> is identified in this image?

Reproducibility. The SFT and GRPO stages support checkpoint resumption. Missing images are skipped during SFT using the `-skip-missing-images` option. Training logs, reward breakdowns, checkpoints, and final model outputs are saved to the specified output and log directories.

C Evaluation Metrics.

We use the same evaluation definitions for QGR, Qwen3-VL, and MedGemma to ensure a controlled comparison. The only difference is the source of the grounding prediction: QGR uses its detector outputs for regions and boxes, whereas Qwen3-VL and MedGemma use the generated <visual_regions> and <visual_locations> fields.

Closed Accuracy. Closed-ended accuracy is computed only on the official closed-question subset:

$$\text{ClosedAcc} = \frac{1}{N_{\text{closed}}} \sum_{i=1}^{N_{\text{closed}}} \mathbb{I}\left(\hat{a}_i = a_i\right) \times 100, \tag{11}$$

where $\hat{a}_i$ and a_i denote the normalized predicted and ground-truth answers, respectively. Single-choice and multi-choice questions are evaluated separately and are not included in `Closed_Accuracy`.

Table 9: MRI question templates.

Open-ended	Closed-ended
1. Does the <LABEL> show enhancement?	1. Does the <LABEL> show enhancement? **Options:** yes, no
2. Is the <LABEL> hyperintense or hypointense?	2. Is the <LABEL> hyperintense or hypointense? **Options:** hyperintense, hypointense, not specified
3. Is the <LABEL> visible in the image?	3. Is the <LABEL> visible in the image? **Options:** yes, no
4. Is there a <LABEL> in the image?	4. Is there a <LABEL> in the image? **Options:** yes, no
5. Is there any additional abnormality besides the recorded finding?	5. Is there any additional abnormality besides the recorded finding? **Options:** yes, no
6. What abnormalities are present in this image?	6. What abnormalities are present in this image?
7. What is the size category of the dominant abnormality?	7. What is the size category of the dominant abnormality? **Options:** small, moderate, large, not specified
8. What signal variation does the <LABEL> show?	8. What signal variation does the <LABEL> show? **Options:** homogeneous, heterogeneous, mixed, not specified
9. Where is the dominant abnormality located?	9. Where is the dominant abnormality located?
10. Which disease or <LABEL> is identified in this image?	10. Which disease or <LABEL> is identified in this image?

Table 10: X-ray question templates.

Open-ended	Closed-ended
1. Is the <LABEL> more opaque or lucent?	1. Is the <LABEL> more opaque or lucent? **Options:** opaque, lucent, not specified
2. Is the <LABEL> sharply marginated or poorly defined?	2. Is the <LABEL> sharply marginated or poorly defined? **Options:** sharply marginated, poorly defined, not specified
3. Is the <LABEL> visible in the image?	3. Is the <LABEL> visible in the image? **Options:** yes, no
4. Is there a <LABEL> in the image?	4. Is there a <LABEL> in the image? **Options:** yes, no
5. Is there any additional abnormality besides the recorded finding?	5. Is there any additional abnormality besides the recorded finding? **Options:** yes, no
6. What abnormalities are present in this image?	6. What abnormalities are present in this image?
7. What is the size category of the dominant abnormality?	7. What is the size category of the dominant abnormality? **Options:** small, moderate, large, not specified
8. What is the size of the <LABEL>?	8. What is the size of the <LABEL>? **Options:** small, moderate, large, not specified
9. Where is the dominant abnormality located?	9. Where is the dominant abnormality located?
10. Which disease or lesion is identified in this image?	10. Which disease or lesion is identified in this image?

Answer Quality. For open-ended questions, we use an LLM judge to score the predicted answer against the reference answer on a scale from 1 to 10. The judge considers helpfulness, relevance, accuracy, and detail, and returns an overall score. The raw answer score is

$$\text{AnswerRaw} = \frac{1}{N_{\text{open}}} \sum_{i=1}^{N_{\text{open}}} s_i^a, \tag{12}$$

Table 11: Ultrasound question templates.

Open-ended	Closed-ended
1. Does the <LABEL> have smooth or irregular margins?	1. Does the <LABEL> have smooth or irregular margins? **Options:** smooth, irregular, not specified
2. Is the <LABEL> homogeneous or heterogeneous?	2. Is the <LABEL> homogeneous or heterogeneous? **Options:** homogeneous, heterogeneous, not specified
3. Is the <LABEL> solid or cystic?	3. Is the <LABEL> solid or cystic? **Options:** solid, cystic, mixed, not specified
4. Is the <LABEL> visible in the image?	4. Is the <LABEL> visible in the image? **Options:** yes, no
5. Is there a <LABEL> in the image?	5. Is there a <LABEL> in the image? **Options:** yes, no
6. Is there any additional abnormality besides the recorded finding?	6. Is there any additional abnormality besides the recorded finding? **Options:** yes, no
7. What abnormalities are present in this image?	7. What abnormalities are present in this image?
8. What is the size category of the dominant abnormality?	8. What is the size category of the dominant abnormality? **Options:** small, moderate, large, not specified
9. Where is the dominant abnormality located?	9. Where is the dominant abnormality located?
10. Which disease or lesion is identified in this image?	10. Which disease or lesion is identified in this image?

where s_i^a is the judge's overall score for the predicted answer.

To reduce judge-scale variation, we additionally normalize each prediction by the score obtained when the reference answer is judged against itself:

$$\text{AnswerNorm} = \frac{100}{N_{\text{open}}} \sum_{i=1}^{N_{\text{open}}} \frac{s_i^a}{\max(s_i^{a,\text{ref}}, 1)}, \tag{13}$$

where $s_i^{a,\text{ref}}$ is the judge score for the reference answer.

Reasoning Quality. Reasoning is evaluated with a separate LLM judge using the same scoring procedure, but with a reasoning-specific prompt. The raw reasoning score is

$$\text{ReasonRaw} = \frac{1}{N_{\text{reason}}} \sum_{i=1}^{N_{\text{reason}}} s_i^r, \tag{14}$$

and the normalized reasoning score is

$$\text{ReasonNorm} = \frac{100}{N_{\text{reason}}} \sum_{i=1}^{N_{\text{reason}}} \frac{s_i^r}{\max(s_i^{r,\text{ref}}, 1)}. \tag{15}$$

Only samples with an available ground-truth reasoning annotation are included.

Region Score. Region grounding is evaluated using strict F1 over normalized region labels. No substring or partial matching is used. For each example,

$$P_i = \frac{|R_i^{\text{pred}} \cap R_i^{\text{gt}}|}{|R_i^{\text{pred}}|}, \qquad R_i = \frac{|R_i^{\text{pred}} \cap R_i^{\text{gt}}|}{|R_i^{\text{gt}}|}, \tag{16}$$

and

$$F1_i = \frac{2P_iR_i}{P_i + R_i}. \tag{17}$$

Table 12: Modality-wise performance on Med4VQA (%). Zero-shot foundation VLMs are compared with SFT and GRPO variants of VLM-R1, MedGemma, Qwen3-VL, and the proposed QGR.

Model	Mod.	Closed	Open	Reason	Region	Box	Final
Other Zero-shot Foundation VLMs							
BLIP2-OPT-2.7B Li et al. (2023b)	CT	8.40	10.92	10.03	0.00	0.00	5.87
	MRI	12.00	10.00	10.00	0.00	0.00	6.40
	US	2.19	10.00	10.40	0.00	0.00	4.52
	X-ray	0.00	10.09	10.02	0.00	0.00	4.02
LLaVA-1.5-7B Liu et al. (2023)	CT	9.24	21.78	13.46	0.00	0.00	8.90
	MRI	12.00	22.48	12.16	0.00	0.00	9.33
	US	7.30	14.96	13.17	0.00	0.00	7.08
	X-ray	7.87	21.18	12.10	0.00	0.00	8.23
LLaVA-Med Li et al. (2023a)	CT	3.36	10.24	10.03	0.00	0.00	4.72
	MRI	20.00	10.00	10.00	0.00	0.00	8.00
	US	7.30	10.00	10.09	0.00	0.00	5.48
	X-ray	10.24	10.03	10.05	0.00	0.00	6.07
QoQ-Med-VL-7B Dai et al. (2026)	CT	13.45	36.03	42.71	42.67	7.38	28.45
	MRI	21.60	35.20	50.16	29.20	11.47	29.53
	US	8.76	30.27	46.61	33.20	12.33	26.23
	X-ray	2.36	35.11	41.95	35.20	6.02	24.13
VLM-R1 Style RL							
VLM-R1-OVD Shen et al. (2025) (Zero-shot)	CT	7.56	30.46	34.13	37.60	6.24	23.20
	MRI	36.00	35.20	38.72	15.20	7.78	26.58
	US	20.44	26.79	35.87	28.40	7.26	23.75
	X-ray	19.69	31.72	37.55	15.20	3.88	21.61
VLM-R1 (SFT)	CT	6.72	29.01	37.91	49.20	7.00	25.97
	MRI	34.40	43.12	47.92	54.80	10.01	38.05
	US	21.17	32.12	45.52	31.20	15.68	29.14
	X-ray	26.77	35.20	37.10	47.60	5.94	30.52
VLM-R1 (GRPO)	CT	6.72	29.24	37.64	49.20	6.80	25.92
	MRI	35.20	42.64	48.32	55.20	10.14	38.30
	US	21.17	32.12	44.20	30.40	15.07	28.59
	X-ray	26.77	36.34	36.42	47.60	5.75	30.58
MedGemma Family							
MedGemma Sellergren et al. (2025) (Zero-shot)	CT	10.08	29.16	24.68	6.00	0.73	14.13
	MRI	8.80	17.68	20.88	0.40	0.87	9.73
	US	4.38	24.16	28.11	1.20	1.19	11.81
	X-ray	2.36	35.36	28.55	1.60	1.49	13.87
MedGemma Sellergren et al. (2025) (SFT)	CT	78.15	53.97	45.82	59.20	2.45	47.92
	MRI	92.80	94.00	66.88	96.40	4.72	70.96
	US	84.67	78.50	73.84	99.20	40.50	75.34
	X-ray	89.76	86.33	68.77	100.00	8.73	70.72
MedGemma Sellergren et al. (2025) (GRPO)	CT	77.31	53.51	45.49	59.20	2.37	47.58
	MRI	89.60	92.56	66.56	95.60	5.04	69.87
	US	84.67	76.50	73.20	99.60	40.08	74.81
	X-ray	90.55	89.01	68.45	100.00	8.95	71.39
Qwen3-VL Family							
Qwen3-VL-4B Yang et al. (2025) (Zero-shot)	CT	14.29	35.74	44.47	28.27	0.88	24.73
	MRI	19.20	41.92	57.28	11.60	0.05	26.01
	US	8.03	31.77	54.23	50.80	2.15	29.39
	X-ray	7.09	45.48	63.03	10.40	7.71	26.74
Qwen3-VL-4B Yang et al. (2025) (SFT)	CT	69.75	62.65	52.59	66.80	8.76	52.11
	MRI	90.40	93.52	69.84	97.20	18.12	73.82
	US	70.80	93.98	76.94	100.00	47.05	77.75
	X-ray	78.74	98.12	74.91	100.00	23.79	75.11
Qwen3-VL-4B Yang et al. (2025) (GRPO)	CT	68.07	63.13	53.41	67.20	8.59	52.08
	MRI	88.80	93.52	69.52	97.20	17.40	73.29
	US	73.72	94.69	76.51	100.00	47.18	78.42
	X-ray	78.74	98.29	74.66	100.00	24.11	75.16
QGR (Ours)							
QGR(MedGemma)-SFT	CT	78.15	58.71	49.19	61.20	1.60	49.77
	MRI	93.60	88.80	67.60	96.80	32.80	75.92
	US	83.94	77.08	72.92	100.00	97.20	86.23
	X-ray	92.13	87.15	72.67	100.00	34.40	77.27
QGR(Qwen3)-SFT	CT	87.39	65.00	55.20	57.47	16.40	56.29
	MRI	93.60	87.68	69.12	95.20	56.40	80.40
	US	84.67	78.32	74.08	100.00	99.20	87.25
	X-ray	94.49	87.80	73.82	100.00	70.00	85.22

The final region score is

$$\text{RegionScore} = \frac{100}{N} \sum_{i=1}^{N} F1_i. \tag{18}$$

For QGR, the predicted regions are obtained from the detector outputs, with generated text used only as a fallback when no detector region is available. For Qwen3-VL and MedGemma, the predicted regions are obtained directly from the generated <visual_regions> field.

Box Score. We use the same Med4-style top-1 localization metric for all models. For each sample, only the first predicted box is used. Its IoU is computed against every ground-truth box, and the best overlap is retained:

$$\mathrm{IoU}_i^{\mathrm{top1}} = \max_{b \in B_i^{\mathrm{gt}}} \mathrm{IoU}\left(\hat{b}_i^{(1)}, b\right). \tag{19}$$

A sample receives a localization score of one if the top-1 IoU exceeds the threshold $\tau = 0.10$:

$$s_i^{\mathrm{box}} = \mathbb{I}\left(\mathrm{IoU}_i^{\mathrm{top1}} \geq 0.10\right). \tag{20}$$

The final box score is

$$\mathrm{BoxScore} = \frac{100}{N} \sum_{i=1}^{N} s_i^{\mathrm{box}}. \tag{21}$$

Thus, Box_score represents top-1 localization accuracy at IoU 0.10 and is not mean IoU.

For QGR, the top-1 box is the highest-ranked detector prediction. GEMeX ground-truth coordinates are converted to the loaded image coordinate system before computing IoU. For Qwen3-VL and MedGemma, no detector is used; the top-1 box is the first box generated inside the <visual_locations> field, with both prediction and ground truth converted to the same evaluation coordinate space.

Final Score. The overall score follows the Med4 evaluation protocol:

$$\mathrm{FinalScore} = \frac{\mathrm{ClosedAcc} + \mathrm{AnswerNorm} + \mathrm{ReasonNorm} + \mathrm{RegionScore} + \mathrm{BoxScore}}{5}. \tag{22}$$

The raw answer and reasoning judge scores are reported for analysis but are not included in the final score.

D Broader Impacts

This work aims to improve trustworthy medical vision–language models by explicitly aligning model reasoning with visual evidence. The primary positive societal impact is improved interpretability in medical VQA systems, where explicit visual grounding helps reasoning remain aligned with disease-relevant image regions rather than producing fluent but weakly grounded explanations. By supervising detector queries using semantic regions and bounding boxes, QGR improves consistency between answer generation and clinical interpretation, reduces hallucinated reasoning, and increases reliability for clinical decision support. This can support clinicians by providing answers that are not only accurate but also visually grounded in localized medical evidence.

However, there are also important risks. Medical VLMs may still produce incorrect or overconfident predictions, and users may over-rely on automated outputs without sufficient clinical verification. Inaccurate localization or hallucinated reasoning could negatively affect diagnostic decisions if deployed without expert supervision. Additionally, dataset bias, annotation noise, and domain shift across hospitals or imaging devices may reduce robustness in real-world settings and may limit generalization across clinical environments.

For these reasons, QGR is intended as a decision-support framework rather than a replacement for clinical expertise. Human verification remains essential, and future work should further study fairness, robustness, uncertainty estimation, and safe deployment in real-world clinical workflows.

E Limitations

Although QGR improves grounded medical reasoning and localization, several limitations remain. First, the framework relies heavily on accurate semantic region annotations and bounding-box supervision for query assignment and localization learning. When annotations are noisy, incomplete, or inconsistent across datasets, detector-query supervision may become unstable, which can negatively

affect both reasoning quality and box localization performance. This is particularly challenging in medical imaging, where expert annotation is expensive and inter-observer variability is common.

Second, QGR depends on the quality of the detector branch. Since detector queries serve as structured reasoning units, weak object proposals or inaccurate bounding boxes from the detection transformer can propagate errors into reasoning generation and region prediction. In such cases, the model may still produce fluent explanations that are partially misaligned with true pathological evidence.

Third, although QGR demonstrates strong performance on Med4VQA and out-of-domain generalization on SLAKE, benchmark performance may not fully reflect real-world clinical deployment. Domain shifts across hospitals, imaging devices, acquisition protocols, and patient populations may reduce robustness. In addition, clinical decision-making often requires longitudinal reasoning, multi-image comparison, and richer patient context beyond single-image VQA settings.

Finally, QGR is designed as a decision-support framework rather than an autonomous diagnostic system. Incorrect localization and over-reliance on automated outputs without expert verification could introduce clinical risks. Human oversight remains essential for safe deployment in medical environments.