# GEMeX-RMCoT: An Enhanced Med-VQA Dataset for Region-Aware Multimodal Chain-of-Thought Reasoning

Bo Liu
The Hong Kong Polytechnic University,
Hong Kong S.A.R., China
bokelvin.liu@connect.polyu.hk

Xiangyu Zhao
The Hong Kong Polytechnic University,
Hong Kong S.A.R., China
xiang-yu.zhao@connect.polyu.hk

Along He
Shenzhen University,
Shenzhen, China
healong@szu.edu.cn

Yidi Chen
West China Hospital of Sichuan University,
Chengdu, China
chenyidi1152@126.com

Huazhu Fu*
IHPC, Agency for Science, Technology and Research,
Singapore
hzf@ieee.org

Xiao-Ming Wu*
The Hong Kong Polytechnic University,
Hong Kong S.A.R., China
xiao-ming.wu@polyu.edu.hk

## Abstract

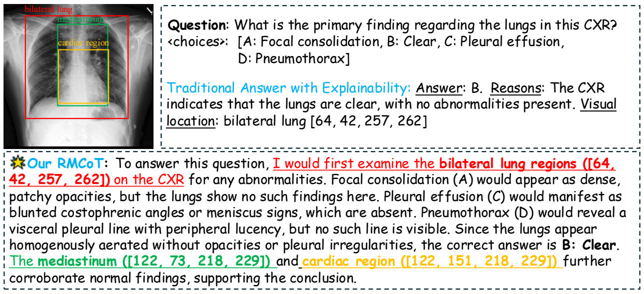
Medical visual question answering aims to support clinical decision-making by enabling models to answer natural language questions based on medical images. While recent advances in multi-modal learning have significantly improved performance, current methods still suffer from limited answer reliability and poor interpretability, impairing the ability of clinicians and patients to understand and trust model outputs. To address these limitations, this work first proposes a **R**egion-Aware **M**ultimodal **C**hain-**o**f-**T**hought (RMCoT) dataset, in which the process of producing an answer is preceded by a sequence of intermediate reasoning steps that explicitly ground relevant visual regions of the medical image, thereby providing fine-grained explainability. Furthermore, we introduce a novel verifiable reward mechanism for reinforcement learning to guide post-training, improving the alignment between the model's reasoning process and its final answer. Remarkably, our method achieves comparable performance using only one-eighth of the training data, demonstrating the efficiency and effectiveness of the proposal. The dataset is available at https://www.med-vqa.com/GEMeX/.

## CCS Concepts

• **Computing methodologies** → **Natural language generation**; **Computer vision tasks**; • **Applied computing** → **Health care information systems**.

## Keywords

Large vision language models, Medical visual question answering, Explainable and trustworthy AI, Chest X-ray dataset



**Question**: What is the primary finding regarding the lungs in this CXR?
<choices>: [A: Focal consolidation, B: Clear, C: Pleural effusion, D: Pneumothorax]

Traditional Answer with Explainability: Answer: B. Reasons: The CXR indicates that the lungs are clear, with no abnormalities present. Visual location: bilateral lung [64, 42, 257, 262]

Our RMCoT: To answer this question, I would first examine the **bilateral lung regions ([64, 42, 257, 262])** on the CXR for any abnormalities. Focal consolidation (A) would appear as dense, patchy opacities, but the lungs show no such findings here. Pleural effusion (C) would manifest as blunted costophrenic angles or meniscus signs, which are absent. Pneumothorax (D) would reveal a visceral pleural line with peripheral lucency, but no such line is visible. Since the lungs appear homogenously aerated without opacities or pleural irregularities, the correct answer is **B: Clear**. The **mediastinum ([122, 73, 218, 229])** and cardiac region ([122, 151, 218, 229]) further corroborate normal findings, supporting the conclusion.

**Figure 1: (Top) Traditional answer with explainability provided by GEMeX: separate textual and visual prompts; (★Bottom) Our multimodal, more detailed thinking process for answer generation, which explicitly grounds evidence in specific regions (shown in the colored texts) of the medical image, i.e., anatomical areas that support diagnosis, thereby enhancing the understanding of questions and answers.**

**ACM Reference Format:**
Bo Liu, Xiangyu Zhao, Along He, Yidi Chen, Huazhu Fu, and Xiao-Ming Wu. 2025. 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 (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3746027.3758277

## 1 Introduction

Medical Visual Question Answering (Med-VQA) has emerged as a promising paradigm for supporting clinical decision-making by enabling models to answer natural language questions based on medical images [17]. Recent advances in multi-modal learning have led to significant performance improvements across a range of Med-VQA benchmarks [7, 9, 14, 20], especially in the era of large vision language models [3, 5, 18, 29, 37, 42]. However, despite these gains, existing methods remain limited in their ability to provide interpretable and trustworthy responses—an essential requirement for real-world deployment in clinical settings. In particular, most current approaches generate answers without revealing the underlying reasoning process or the specific image regions that inform

---

*Corresponding authors.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
*MM '25, Dublin, Ireland.*
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3758277

their decisions. This lack of transparency undermines user trust and hinders the integration of such systems into clinical workflows.

Recently, GEMeX [22] introduced a new dataset where each VQA triplet is paired with a textual reason and a corresponding visual region as an explanation, bridging this gap to some extent. However, the textual and visual prompts are independent and lack coherence. Moreover, the textual reason merely serves as an explanation of the answer rather than providing guidance on how to solve the question based on the image. Therefore, based on it, we propose a new dataset **GEMeX-RMCoT** that enhances interpretability through introducing a **R**egion-Aware **M**ultimodal **C**hain-**o**f-**T**hought reasoning mechanism, as shown in Figure 1. Different from the traditional CoT [38] and recent approaches [16, 25] that rely solely on reinforcement learning to incentivize the thinking process, our RMCoT explicitly decomposes answer generation into intermediate steps, along with grounding corresponding visual evidence within the medical image. By linking reasoning steps to spatially localized regions [32], we offer clinicians and patients a fine-grained explanation of how the model arrives at its conclusions.

While the GEMeX-RMCoT dataset enables the modeling of explicit thinking, our empirical findings indicate supervised fine-tuning (SFT) alone is insufficient to fully unleash the reasoning capabilities of the large vision language models (LVLMs). To address this, we employ reinforcement learning for post-training that builds upon the SFT-tuned model. It introduces a novel verifiable reward mechanism to incentivize alignment between the reasoning trajectory and relevant visual and textual cues, thereby enhancing both the accuracy of intermediate thinking steps and final answer generation. In contrast to existing methods requiring large-scale datasets, our approach is data-efficient, achieving comparable performance using only one-eighth of the full training data.

In summary, our contributions are threefold:

- We introduce GEMeX-RMCoT, a region-aware multimodal chain-of-thought dataset for Med-VQA, which provides step-by-step reasoning explicitly grounded in specific anatomical regions of medical images, offering better interpretability for understanding the questions and answers.
- We perform supervised fine-tuning on the LVLM using our GEMeX-RMCoT dataset, and further apply reinforcement learning with a novel verifiable reward mechanism to incentivize reasoning abilities, enabling the generation of more accurate thinking paths and final answers.
- From comprehensive experiments, we demonstrate the effectiveness of the proposed dataset and reward mechanism that empower LVLM to achieve comparable performance with only one-eighth of the training data.

## 2 Related Work

### 2.1 Medical VQA Dataset

Recent years have witnessed significant progress in medical visual question answering through the development of specialized datasets targeting diverse clinical challenges. VQA-RAD [17] established the foundation with 3,000 question-answer pairs focused on radiology images. SLAKE [21] expanded the scope with over 14,000 manually annotated QA pairs spanning CT, MRI, and X-ray modalities. OmniMedVQA [9] further broadened coverage across multiple body regions and imaging modalities to enhance model generalization capabilities. For specialized tasks, PathVQA [7] offers 32,000 QAs on histopathology images. MIMIC-Diff-VQA [8] emphasizes differential diagnosis between paired X-rays. GEMeX [22] enhances VQA with both textual and visual explanations, facilitating a better understanding of the answers. However, none of them incorporates a reasoning process for problem-solving, which hinders patients and junior doctors from fully understanding the questions and answers. To fill this gap, we propose a region-aware multimodal CoT reasoning dataset, GEMeX-RMCoT, in this work.

### 2.2 Reinforcement Learning

Recently, the growing popularity of reasoning-focused large language models such as GPT-o1 [11] and DeepSeek-R1 [6] has drawn increasing attention to the underlying mechanisms of reinforcement learning (RL). This interest has led to remarkable advancements in tasks such as mathematics [24, 35], code generation [13, 36], hallucination detection [33, 40], and interdisciplinary research [39, 41]. A notable contribution is the Group Relative Policy Optimization (GRPO) algorithm proposed in DeepSeekMath [30], which significantly simplifies the training process. Unlike traditional RL algorithms like PPO [13] that require a critic model to evaluate policy performance, GRPO compares groups of candidate responses directly, eliminating the need for an additional critic. GRPO's simplicity has inspired further exploration into the reasoning capabilities of LVLMs for multimodal and visual tasks [10, 23, 31]. In the medical domain, GRPO has also been applied to incentivize the reasoning abilities of LVLMs for tasks such as out-of-distribution detection [25] and visual question answering [16]. Unlike these works that merely elicit textual descriptions of visual content, we propose a visually grounded reasoning paradigm that enhances interpretability and strengthens evidential support.

## 3 Construction of GEMeX-RMCoT

### 3.1 Task Definition

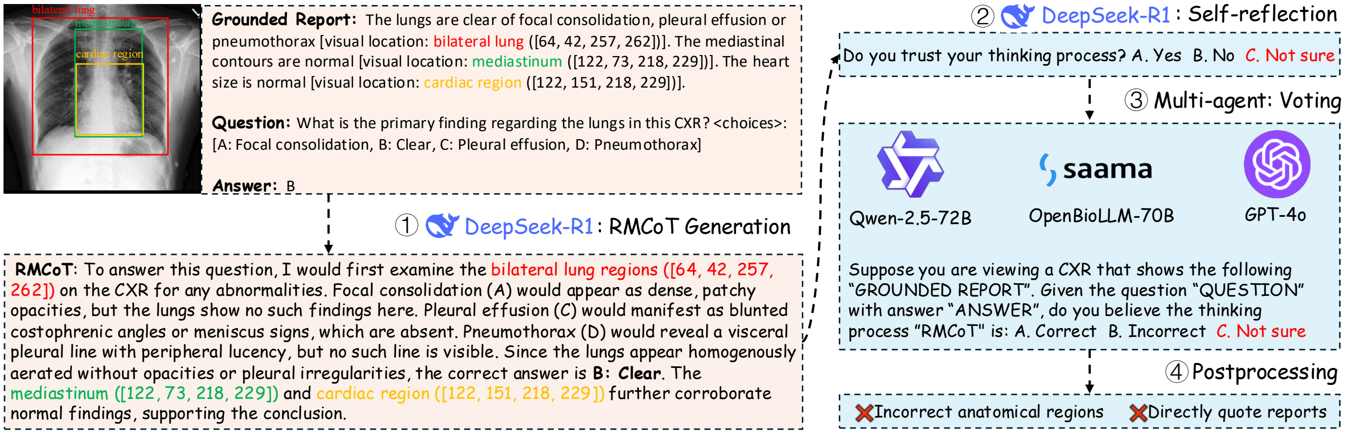
Unlike conventional chain-of-thought approaches used in general tasks such as VQA [34], which typically generate only textual rationales before outputting the final answer, we propose a multimodal-level thinking process with visual grounding tailored for Med-VQA. As illustrated in Figure 1, it decomposes the reasoning process into several steps with visual attention to specific image regions, aiming to deliver more detailed explanations and improved interpretability for patients and junior doctors. Formally, given a question $q$ for image $v$, we hope to provide a multi-modal thinking process $T$ involving visual grounding to attain answer $a$ via

$$T = \left(\{r_i\}_{i=1}^{N_r}, \{t_j\}_{j=1}^{N_t}\right), \tag{1}$$

where $\{r_i\}_{i=1}^{N_r}$ is a set of image regions and $\{t_j\}_{j=1}^{N_t}$ is a set of relevant textual descriptions.

### 3.2 An Uncertainty-Driven Multi-Agent Framework

As illustrated in Figure 2, we construct the RMCoT dataset based on GEMeX [22], the largest Med-VQA dataset that provides detailed visual and textual explanations for each VQA triplet. To ensure the



**Grounded Report:** The lungs are clear of focal consolidation, pleural effusion or pneumothorax [visual location: bilateral lung ([64, 42, 257, 262])]. The mediastinal contours are normal [visual location: mediastinum ([122, 73, 218, 229])]. The heart size is normal [visual location: cardiac region ([122, 151, 218, 229])].

**Question:** What is the primary finding regarding the lungs in this CXR? <choices>: [A: Focal consolidation, B: Clear, C: Pleural effusion, D: Pneumothorax]

**Answer:** B

**RMCoT**: To answer this question, I would first examine the bilateral lung regions ([64, 42, 257, 262]) on the CXR for any abnormalities. Focal consolidation (A) would appear as dense, patchy opacities, but the lungs show no such findings here. Pleural effusion (C) would manifest as blunted costophrenic angles or meniscus signs, which are absent. Pneumothorax (D) would reveal a visceral pleural line with peripheral lucency, but no such line is visible. Since the lungs appear homogenously aerated without opacities or pleural irregularities, the correct answer is **B: Clear**. The mediastinum ([122, 73, 218, 229]) and cardiac region ([122, 151, 218, 229]) further corroborate normal findings, supporting the conclusion.

Do you trust your thinking process? A. Yes B. No C. Not sure

Suppose you are viewing a CXR that shows the following "GROUNDED REPORT". Given the question "QUESTION" with answer "ANSWER", do you believe the thinking process "RMCoT" is: A. Correct B. Incorrect C. Not sure

**Figure 2: Illustration of the proposed uncertainty-driven multi-agent framework for generating the RMCoT dataset. The left (orange) section depicts the data generation process, while the right (blue) section shows the uncertainty-driven quality assurance pipeline, i.e., the “not sure” option prompts a deeper reflection and more deliberate decision-making.**

**messages** = [“role”:“**user**”, “content”:
f“‘Suppose you are viewing a CXR that shows the following: “The hilar contours are normal [visual location: bilateral hilar structures ([116, 112, 227, 182])] ...”. Given the question: “YOUR QUESTION”, provide a detailed thinking process (around 100 words) including specific visual location (e.g., (region [x1,y1,x2,y2])) about how to solve this question with answer “ANSWER”. You must assume that you are viewing the CXR image rather than reading the textual findings, thus do not output words like “observe the report” or “from report” or “report states” or “given findings” or “provided findings” or “described findings”.""

**Table 1: Given a question-answer pair, our tailored prompt guides DeepSeek-R1 to generate RMCoT data.**

high quality of RMCoT, we introduce an uncertainty-driven multi-agent framework, centered around DeepSeek-R1 [6], which also coordinates multiple agents and leverages an uncertainty mechanism to enhance the accuracy of the generated data.

*3.2.1 Data Generation.* In Stage ①, we primarily use DeepSeek-R1 to generate the thinking process for each VQA triplet due to its outstanding reasoning capabilities. However, as DeepSeek-R1 is a text-based LLM and lacks the ability to directly perceive visual inputs, we utilize the grounded reports from GEMeX as an intermediary. In GEMeX’s reports, each sentence is aligned with a specific anatomical region, enabling DeepSeek-R1 to simulate a multimodal reasoning process required for solving the tasks. To optimize the prompt design, we randomly sample 50 questions from each question type (e.g., open-ended) in the training set—resulting in 200 questions in total—to evaluate performance and guide prompt iteration. The final optimized prompt is presented in Table 1.

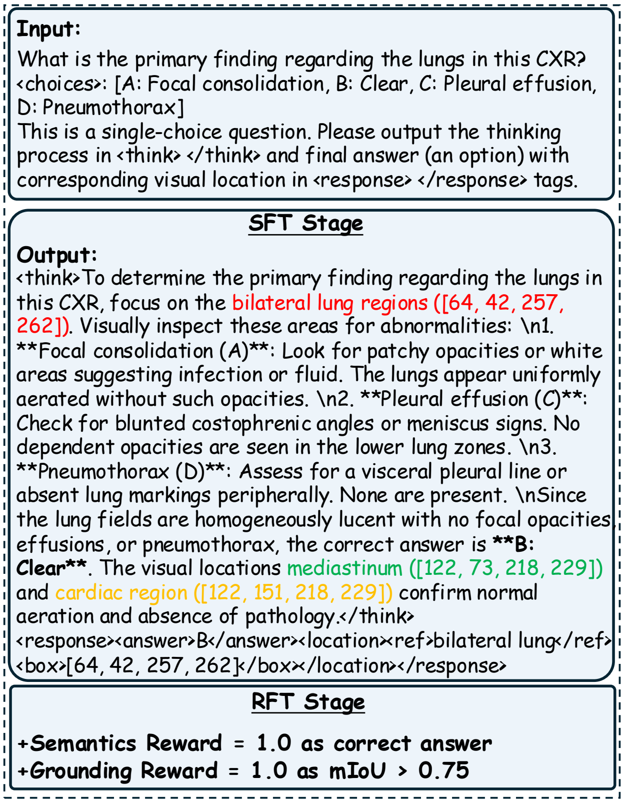
*3.2.2 Uncertainty-driven Quality Assurance.* Although manual verification and prompt modification are applied, the generated data still suffers from quality issues. To address this, we design a pipeline which contains three steps (Stage ② to Stage ④) to ensure quality: (1) We first introduce an uncertainty-aware mechanism that enables DeepSeek-R1 to self-reflect and evaluate the reliability of its own outputs. Specifically, after generating the RMCoT response, we prompt DeepSeek-R1 with the question: *“Do you trust your thinking process? A. Yes B. No C. Not sure”*. This self-reflection with uncertainty choice “Not sure” allows the model to carefully reassess its previous reasoning, which has been shown to improve output accuracy [2]. We retain only those RMCoT responses for which the model answers “A. Yes.”; (2) To avoid bias caused by relying on a single LLM for both data generation and evaluation [26], we introduce additional agents to assess the quality of the generated data. As shown in Stage ③, we employ three additional agents—Qwen-2.5-72B [27], OpenBioLLM-70B[1], and GPT-4o [1]—as external reviewers. These agents are given grounded reports, questions, and answers to assess the accuracy of the thinking process. Similar to the self-reflection stage, we introduce an “uncertain” option to provide the agents with greater flexibility for consideration and decision-making. This process helps eliminate incorrect reasoning paths generated by DeepSeek-R1 based on its internal knowledge. Then, we adopt a voting mechanism: the RMCoT is retained only if all three agents unanimously select “A. Correct” option; (3) Finally, we apply post-processing to eliminate RMCoT outputs with significant errors, such as incorrect location information or instances that quote the report or findings directly, rather than generating content from an observational perspective.

## 3.3 Radiologist Validation

After auto-generation, a radiologist randomly samples 200 examples from each of the four question types in the dataset for manual inspection. Cross-validation reveals no clinically inconsistent reasoning chains, demonstrating the effectiveness of the proposed framework. In total, we generate 202,384 RMCoT instances for 21,994 images, all drawn from the GEMeX training set. The detailed distribution is presented in Table 2.

---

[1] https://huggingface.co/aaditya/Llama3-OpenBioLLM-70B



**Input:**
What is the primary finding regarding the lungs in this CXR?
<choices>: [A: Focal consolidation, B: Clear, C: Pleural effusion, D: Pneumothorax]
This is a single-choice question. Please output the thinking process in <think> </think> and final answer (an option) with corresponding visual location in <response> </response> tags.

**SFT Stage**

**Output:**
<think>To determine the primary finding regarding the lungs in this CXR, focus on the bilateral lung regions ([64, 42, 257, 262]). Visually inspect these areas for abnormalities: \n1. **Focal consolidation (A)**: Look for patchy opacities or white areas suggesting infection or fluid. The lungs appear uniformly aerated without such opacities. \n2. **Pleural effusion (C)**: Check for blunted costophrenic angles or meniscus signs. No dependent opacities are seen in the lower lung zones. \n3. **Pneumothorax (D)**: Assess for a visceral pleural line or absent lung markings peripherally. None are present. \nSince the lung fields are homogeneously lucent with no focal opacities, effusions, or pneumothorax, the correct answer is **B: Clear**. The visual locations mediastinum ([122, 73, 218, 229]) and cardiac region ([122, 151, 218, 229]) confirm normal aeration and absence of pathology.</think>
<response><answer>B</answer><location><ref>bilateral lung</ref><box>[64, 42, 257, 262]</box></location></response>

**RFT Stage**

**+Semantics Reward = 1.0 as correct answer**
**+Grounding Reward = 1.0 as mIoU > 0.75**

**Figure 3: An example from the SFT and RFT stages: after SFT, the model learns to think carefully and incorporates grounding to specific pathological regions for visual evidence. During the RFT stage, since both the answer and the grounded regions are correct, the two reward scores are each 1.0.**

| Open-ended | Closed-ended | Single-choice | Multi-choice |
|---|---|---|---|
| 61,240 | 36,432 | 52,850 | 51,862 |

**Table 2: Distribution of generated RMCoT data across the four question types in GEMeX.**

# 4 Fine-tuning on GEMeX-RMCoT: Learning to Think with Visual Grounding in Med-VQA

## 4.1 Supervised Fine-tuning

After generating the RMCoT data, we first train the large vision language model (LVLM) using Supervised Fine-tuning (SFT) to equip it with task-aware thinking capabilities. To generate the outputs required by GEMeX—which include both the final answer and its corresponding visual location (omitting the reasons, as we provide a more detailed preceding thinking path)—the model must produce structured outputs in two distinct parts. First, it should output a thinking trace that incorporates explicit visual grounding, enclosed within the `<think>...</think>` tags. Second, it should deliver a concise final response that includes the predicted "answer" and the summarized "visual location", wrapped within a designated XML-like structure: `<response>` `<answer>...</answer>` `<location>...</location></response>`. This format ensures both interpretability and alignment with the GEMeX evaluation requirements. An illustrative example is shown in Figure 3.

**messages** = [“role”:“**user**”, “content”: f“‘We would like to request your feedback on the performance of two AI assistants in response to the user question displayed above. For your reference, the visual content in the image is represented with caption describing the same image. Please rate the accuracy (most important), relevance of their responses, considering both answer and reason (if any). Each assistant receives an overall score on a scale of 1 to 10, where a higher score indicates better overall performance. Please output both scores and your reaon in JSON format { assistant1: score, assistant2: score, reason:your_reason }."""

**Table 3: Our designed prompt for *Semantics Reward.***

## 4.2 Reinforcement Fine-tuning

Although SFT can significantly improve thinking capabilities, we have observed in practice that: (1) due to the complexity of the task, the model sometimes generates incorrect thinking process after SFT, for example, incorrect analysis of images, which may be because SFT is better at helping LVLMs memorize rather than truly understand [4]; (2) inaccurate region localization within the thinking path can propagate errors to the final answer, ultimately leading to degraded performance. Therefore, we adopt Reinforcement Fine-tuning (RFT), specifically through Group Relative Policy Optimization (GRPO) [30], as a post-training strategy to further activate and enhance the model's thinking capabilities with a newly proposed reward mechanism.

*4.2.1 Group Relative Policy Optimization.* Given a question $q$ from question set $Q$, GRPO first samples a set of outputs $\{o_i\}_i^G$ from the policy model $\pi_{\theta_{old}}$ during each learning iteration. Then it optimizes the LVLM $\pi_\theta$ by maximizing the following objective:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim P(Q),\,\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(O|q)}\Bigg[ \frac{1}{G}\sum_{i=1}^{G}\min\left(\frac{\pi_\theta(o_i\mid q)}{\pi_{\theta_{\text{old}}}(o_i\mid q)}A_i,\text{clip}\left(\frac{\pi_\theta(o_i\mid q)}{\pi_{\theta_{\text{old}}}(o_i\mid q)},1-\varepsilon,1+\varepsilon\right)A_i\right) - \beta D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})\Bigg] \tag{2}$$

where $A_i = \frac{r_i-\text{mean}(\{r_1,r_2,\ldots,r_G\})}{\text{std}(\{r_1,r_2,\ldots,r_G\})}$ is the relative advantage over group rewards $\{r_i\}_i^G$, $D_{\text{KL}}$ is the KL divergence between $\pi_\theta$ and $\pi_{\text{ref}}$ (the model after SFT stage) to prevent catastrophic forgetting caused by over optimization, $\epsilon$ and $\beta$ are the PPO clipping hyperparameter [28] and the coefficient for the KL penalty, respectively.

*4.2.2 Reward Functions.* We propose two reward functions, tailored for the two output parts of GEMeX: answer and involved visual locations. (1) *Semantics Reward*: In GEMeX, four types of questions are defined (Table 2). For choice-based and closed-ended questions, reward design is straightforward, as we can directly use Accuracy Reward [25] to evaluate whether the model's outputs are correct. Nevertheless, designing a reward function for open-ended questions is particularly challenging due to the absence of fixed

| | Training Paradigm | | | Open-ended | | Closed-ended | | Single-choice | | Multi-choice | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Dataset | SFT | RFT | A-score† | V-score | A-score | V-score | A-score | V-score | A-score | V-score | A-score.Avg |
| Zero-shot | - | | | 85.13 | 12.31 | 30.39 | 20.23 | 49.00 | 27.81 | 12.02 | 7.59 | 44.14 |
| Fine-tuning | GEMeX-Full | ✓ | - | **97.31** | **60.51** | 82.32 | **69.32** | 83.08 | **67.38** | **74.51** | **55.98** | 84.31 |
| | GEMeX-200K | ✓ | - | 92.54 | 56.59 | 79.93 | 66.04 | 78.92 | 61.39 | 67.93 | 49.71 | 79.83 |
| | GEMeX-RMCoT-200K | ✓ | ✗ | 94.71 | 56.52 | 87.11 | 66.23 | 80.08 | 61.76 | 70.20 | 51.13 | 83.03 |
| | GEMeX-RMCoT-200K | ✓ | ✓ | 96.44 | 59.63 | **90.24** | 68.39 | **83.31** | 64.52 | 73.18 | 54.49 | **85.79** |

**Table 4: Performance of Qwen2.5-VL-7B with different evaluating paradigms on GEMeX. The A-score indicates answer or choice accuracy (%), and the V-score represents mIoU (%). † indicates that the accuracy of the answer is judged by the LLM, as the question is open-ended. The best results are bolded, and the second-best are underlined in each column.**

answers. Unlike [19] which employs traditional machine learning metrics such as BLEU as rewards—metrics that have been shown to poorly capture essential medical content [12]—we introduce a semantics-aware accuracy reward. It enables the integration of all four types into a unified, coherent reward paradigm. Specifically, we input both the ground truth (GT) answer and the model-generated answer into OpenBioLLM-70B, which can provide two accuracy scores at the semantics level. When these two scores are close, i.e., if the absolute difference between them is less than 2, we assign a reward of 1 point; otherwise, the reward is 0. The specific prompt is shown in Table 3. (2) *Grounding Reward*: Besides giving answers, the model needs to provide the visual locations involved in solving the questions. Therefore, we first check whether the number of bounding boxes predicted by the model matches the number in the GT. If they match, we then evaluate whether the mean Intersection over Union (mIoU) exceeds 0.75. A reward score of 1.0 is assigned only when both conditions are satisfied; otherwise, the reward is 0.

# 5 Experiments

## 5.1 Dataset

We conduct experiments on the GEMeX dataset and randomly sample a portion of its training set to generate RMCoT data. We denote the version with the complete training set of 1.59M VQA triples as GEMeX-Full. The 200K subset that includes RMCoT data is denoted as GEMeX-RMCoT-200K, while the subset without RMCoT is referred to as GEMeX-200K. For evaluation, we use the original GEMeX test set, which consists of 300 images and 3,960 questions (1,144 open-ended questions, 543 closed-ended questions, 1,300 single-choice questions, and 973 multiple-choice questions).

## 5.2 Training Details

We mainly explore the Qwen2.5-VL-7B-Instruct for experimental verification. During the SFT stage, we fine-tune both the visual projection layers and the LLM components. Particularly, the model is trained for 2 epochs on 8 NVIDIA H100 GPUs with a batch size of 256. The network is warmed up in the first 0.05 epochs with a linear learning rate from 3e-7 to 1e-4, which further decays by cosine schedule. The optimizer is AdamW. During the RFT stage, the model is trained for 1 epoch on 8 NVIDIA H100 GPUs with a batch size of 128. The learning rate is $2e-6$. Regarding the hyper-parameters in Eq. 2, we generate 8 rollouts (i.e., $G$) for each input and set $\beta$ and $\epsilon$ to $1e-3$ and 0.2, respectively. The sample generator and LLM used for semantics reward are deployed by vLLM [15].

| Task | SFT Data | Original A-score | Per.1 | Per.2 | Per.3 | Std.† |
|---|---|---|---|---|---|---|
| Closed. | 200K | 79.93 | 76.98 | 83.06 | 75.69 | 3.214 |
| | RMCoT-200K | 87.11 | 88.39 | 86.56 | 87.66 | **0.752** |
| Single. | 200K | 78.92 | 79.00 | 77.69 | 77.31 | 0.723 |
| | RMCoT-200K | 80.08 | 80.15 | 80.00 | 79.69 | **0.192** |

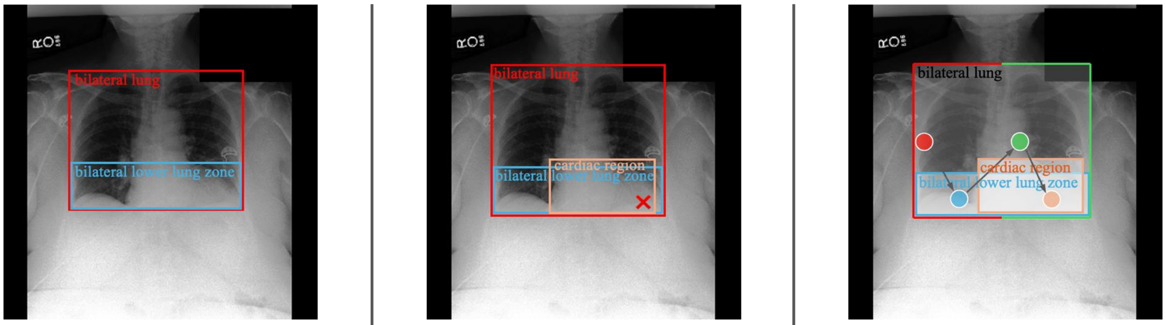
**Table 5: Model performance under different perturbations. † means the standard deviation across the three perturbations from original A-score.**

## 5.3 Main Results

We report the main results in Table 4. It can be seen that (1) Compared to zero-shot learning, fine-tuning can significantly improve the model's performance; (2) When SFT is conducted with only 200K vanilla data (i.e., GEMeX-200K), the model lags behind the one trained on the full dataset by a large margin, with an average A-score that is around 4.5% lower; (3) When RMCoT data (GEMeX-RMCoT-200K) is further used for SFT, the model's performance improves substantially, even surpassing that of the model trained on the full dataset in certain tasks, e.g., an improvement of approximately **5%** on closed-ended tasks. Nevertheless, it can be observed that when only SFT is used, the improvement in visual grounding is quite limited. This is often due to inaccurate location reasoning in the thinking path after learning RMCoT with SFT, which leads to error propagation in the final outputs; (4) To address this issue, further applying RFT with semantics reward and grounding reward yields noticeable enhancement, a **1.5%–3%** improvement in both answer accuracy and localization performance. It is worth noting that, compared to using the full training data, the model further fine-tuned with RFT achieves better results in answer generation (i.e., higher average A-score) and comparable performance in localization. These findings demonstrate the effectiveness of our proposed RMCoT dataset and RFT rewards.

## 5.4 Robustness to Input Variations

We conduct two common perturbations to generate input variations: (1) we modify closed-ended questions while preserving their semantic integrity—for example, changing "Is the heart size


bilateral lung
bilateral lower lung zone
bilateral lung
cardiac region
bilateral lower lung zone
bilateral lung
cardiac region
bilateral lower lung zone

**Question**: Which findings are noted in this CXR? choices: [A: Low lung volumes, B: Atelectasis, C: Pleural effusions, D: Cardiomegaly]

**Answer**: [A, B]

**Reason**: The CXR notes decreased lung volumes and the presence of mild atelectasis, while pleural effusions and cardiomegaly are absent.

**Visual Region**: [bilateral lung, bilateral lower lung zone]

**Visual Location**: [[73, 65, 254, 223], [73, 168, 254, 223]]

Ground Truth from GEMeX

**<response>**
<answer>[A, B, D] ✗ </answer>
<reason>
The CXR shows low lung volumes, atelectasis, and mild cardiomegaly ✗.
</reason>
<location>
<ref>bilateral lung</ref>
<box>[70, 65, 259, 227]</box>
<ref>bilateral lower lung zone</ref>
<box>[70, 168, 259, 227]</box>
<ref>cardiac region ✗ </ref>
<box>[134, 158, 239, 226]</box>
</location>
**</response>**

SFT w/o RMCoT

**<think>**
To solve this question, analyze the CXR image systematically by focusing on the specified visual locations: **1. **Low lung volumes**: Check the bilateral lung regions ([73, 65, 259, 232]). The lungs appear smaller than expected, with crowded vascular markings and elevated diaphragms, indicating reduced volume.** **2. **Atelectasis**: Inspect the bilateral lower lung zones ([73, 164, 259, 232]). There are subtle linear or wedge-shaped opacities at the bases, consistent with mild atelectasis.** **3. **Pleural effusions**: Re-examine the bilateral lung regions ([73, 65, 259, 232]). No blunting of costophrenic angles or meniscus signs are seen, ruling out effusions.** **4. **Cardiomegaly**: Assess the cardiac region ([133, 158, 244, 229]). The cardiothoracic ratio is normal (\textless50\%), excluding cardiomegaly.** Thus, the correct findings are **A (Low lung volumes)** and **B (Atelectasis)**.
**</think>**
**<response>**
<answer>[A, B]</answer>
<location>
<ref>bilateral lung</ref>
<box>[73, 65, 259, 232]</box>
<ref>bilateral lower lung zone</ref>
<box>[73, 164, 259, 232]</box>
</location>
**</response>** ✓

RFT w/ RMCoT

**Figure 4: One challenging example from GEMeX answered by models trained with and without RMCoT. (✓) or (✗) in outputs highlight correct or incorrect reasons or answers. The colored words indicate the thinking with visual grounding process.**

abnormal in this CXR?" by randomly replacing the interrogative word "is" with "isn't", or replacing "abnormal" with "normal", and adjusting the corresponding answer accordingly. During the test phase, each question has a 50% chance of being modified by one of the two modes mentioned above. If a question does not meet the criteria for either mode, no modification will be applied; (2) we randomly change the order of the options in single-choice questions. For each type of perturbation, we conduct three rounds of test and show models' performance in Table 5. To more clearly compare the effectiveness of RMCoT, we compare the performance of the models only after SFT. The results show that after fine-tuning the model with RMCoT, its robustness against interference is significantly improved (i.e., lower standard deviation), highlighting another advantage, especially in the context of potential challenges encountered in practical deployment.

## 5.5 Qualitative Evaluation

As shown in Figure 4, we compare the performance of models trained with and without RMCoT on GEMeX-200K. When using RMCoT, we present the model outputs after RFT stage. It can be seen that when using RMCoT with proper tuning, not only are the answers more accurate, but the model also carefully thinks with analyzing corresponding regions in the image when addressing questions, leading to stronger overall interpretability. More examples are shown on the project page.

## 6 Conclusion

In this work, we address the critical issue of interpretability in Medical Visual Question Answering by introducing a region-aware multimodal chain-of-thought dataset (RMCoT). It explicitly links thinking steps to corresponding regions in the medical image, offering fine-grained visual explanations that enhance transparency and trust in model-generated answers. To further guide the model in generating accurate thinking paths, we proposed a new reward mechanism for reinforcement learning that aligns the model's outputs with relevant visual and textual evidence. Despite using a subset of training data, the trained model is able to achieve impressive performance. We believe this work marks a step toward making Med-VQA systems more explainable and clinically usable, paving the way for safer and more reliable AI-assisted diagnosis.

## Acknowledgments

We thank all reviewers for their valuable time and feedback. This work was supported by Project P0056021 under Parent Project P0050643 of the Otto Poon Charitable Foundation Smart Cities Research Institute, HK PolyU; the Research Student Attachment Programme, HK PolyU; and the A*STAR Central Research Fund ("Robust and Trustworthy AI system for Multi-modality Healthcare"), Singapore.

## References

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

[2] Jiuhai Chen and Jonas Mueller. 2023. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. *arXiv preprint arXiv:2308.16175* (2023).

[3] Xiaolan Chen, Ziwei Zhao, Weiyi Zhang, Pusheng Xu, Le Gao, Mingpu Xu, Yue Wu, Yinwen Li, Danli Shi, and Mingguang He. 2024. EyeGPT: Ophthalmic Assistant with Large Language Models. *arXiv preprint arXiv:2403.00840* (2024).

[4] Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. 2025. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161* (2025).

[5] Jiaqi Cui, Lu Wen, Yuchen Fei, Bo Liu, Luping Zhou, Dinggang Shen, and Yan Wang. 2025. HiLa: Hierarchical Vision-Language Collaboration for Cancer Survival Prediction. *arXiv preprint arXiv:2507.04613* (2025).

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

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

[8] Xinyue Hu, Lin Gu, Qiyuan An, Mengliang Zhang, Liangchen Liu, Kazuma Kobayashi, Tatsuya Harada, Ronald M Summers, and Yingying Zhu. 2023. Expert knowledge-aware image difference graph representation learning for difference-aware medical visual question answering. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4156–4165.

[9] Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. 2024. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22170–22183.

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

[11] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720* (2024).

[12] Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463* (2021).

[13] Fangkai Jiao, Geyang Guo, Xingxing Zhang, Nancy F Chen, Shafiq Joty, and Furu Wei. 2024. Preference Optimization for Reasoning with Pseudo Feedback. *arXiv preprint arXiv:2411.16345* (2024).

[14] Yash Khare, Viraj Bagal, Minesh Mathew, Adithi Devi, U Deva Priyakumar, and CV Jawahar. 2021. Mmbert: Multimodal bert pretraining for improved medical vqa. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 1033–1036.

[15] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*. 611–626.

[16] Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, and Xiaofeng Yang. 2025. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *arXiv preprint arXiv:2503.13939* (2025).

[17] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data* 5, 1 (2018), 1–10.

[18] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems* 36 (2024).

[19] Xu Liang. 2025. Group Relative Policy Optimization for Image Captioning. *arXiv preprint arXiv:2503.01333* (2025).

[20] Bo Liu, Li-Ming Zhan, and Xiao-Ming Wu. 2021. Contrastive pre-training and representation distillation for medical visual question answering based on radiology images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*. Springer, 210–220.

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

[22] Bo Liu, Ke Zou, Liming Zhan, Zexin Lu, Xiaoyu Dong, Yidi Chen, Chengqiang Xie, Jiannong Cao, Xiao-Ming Wu, and Huazhu Fu. 2024. GEMeX: A Large-Scale, Groundable, and Explainable Medical VQA Benchmark for Chest X-ray Diagnosis. *arXiv preprint arXiv:2411.16778* (2024).

[23] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. 2025. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785* (2025).

[24] Trung Quoc Luong, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. 2024. Reft: Reasoning with reinforced fine-tuning. *arXiv preprint arXiv:2401.08967* 3 (2024).

[25] Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. 2025. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. *arXiv preprint arXiv:2502.19634* (2025).

[26] Arjun Panickssery, Samuel Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems* 37 (2024), 68772–68802.

[27] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 Technical Report. arXiv:2412.15115 [cs.CL]

[28] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[29] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, Lu Yang, Kejia Chen, Per Bjornsson, Shashir Reddy, Ryan Brush, Kenneth Philbrick, Mercy Asiedu, Ines Mezerreg, Howard Hu, Howard Yang, Richa Tiwari, Sunny Jansen, Preeti Singh, Yun Liu, Shekoofeh Azizi, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Riviere, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Elena Buchatskaya, Jean-Baptiste Alayrac, Dmitry Lepikhin, Vlad Feinberg, Sebastian Borgeaud, Alek Andreev, Cassidy Hardin, Robert Dadashi, Léonard Hussenot, Armand Joulin, Olivier Bachem, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Clement Farabet, Joelle Barral, Tris Warkentin, Jonathon Shlens, David Fleet, Victor Cotruta, Omar Sanseviero, Gus Martins, Phoebe Kirk, Anand Rao, Shravya Shetty, David F. Steiner, Can Kirmizibayrak, Rory Pilgrim, Daniel Golden, and Lin Yang. 2025. MedGemma Technical Report. arXiv:2507.05201 [cs.AI] https://arxiv.org/abs/2507.05201

[30] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300* (2024).

[31] Zhaochen Su, Linjie Li, Mingyang Song, Yunzhuo Hao, Zhengyuan Yang, Jun Zhang, Guanjie Chen, Jiawei Gu, Juntao Li, Xiaoye Qu, et al. 2025. Openthinkimg: Learning to think with images via visual tool reinforcement learning. *arXiv preprint arXiv:2505.08617* (2025).

[32] Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, Linjie Li, Yu Cheng, Heng Ji, Junxian He, and Yi R. Fung. 2025. Thinking with Images for Multimodal Reasoning: Foundations, Methods, and Future Frontiers. arXiv:2506.23918 [cs.CV]

[33] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525* (2023).

[34] Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. 2025. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605* (2025).

[35] Huaiyuan Ying, Shuo Zhang, Linyang Li, Zhejian Zhou, Yunfan Shao, Zhaoye Fei, Yichuan Ma, Jiawei Hong, Kuikun Liu, Ziyi Wang, et al. 2024. Internlm-math:

Open math large language models toward verifiable reasoning. *arXiv preprint arXiv:2402.06332* (2024).

[36] Kechi Zhang, Ge Li, Yihong Dong, Jingjing Xu, Jun Zhang, Jing Su, Yongfei Liu, and Zhi Jin. 2024. Codedpo: Aligning code models with self generated and verified source code. *arXiv preprint arXiv:2410.05605* (2024).

[37] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. 2023. Biomed-CLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915* (2023).

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

[39] Xiangyu Zhao, Wanghan Xu, Bo Liu, Yuhao Zhou, Fenghua Ling, Ben Fei, Xiaoyu Yue, Lei Bai, Wenlong Zhang, and Xiao-Ming Wu. 2025. MSEarth: A Benchmark for Multimodal Scientific Comprehension of Earth Science. *arXiv preprint arXiv:2505.20740* (2025).

[40] Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. 2023. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839* (2023).

[41] Yuhao Zhou, Yiheng Wang, Xuming He, Ruoyao Xiao, Zhiwei Li, Qiantai Feng, Zijie Guo, Yuejin Yang, Hao Wu, Wenxuan Huang, et al. 2025. Scientists' First Exam: Probing Cognitive Abilities of MLLM via Perception, Understanding, and Reasoning. *arXiv preprint arXiv:2506.10521* (2025).

[42] Ke Zou, Yang Bai, Bo Liu, Yidi Chen, Zhihao Chen, Yang Zhou, Xuedong Yuan, Meng Wang, Xiaojing Shen, Xiaochun Cao, et al. 2025. Uncertainty-aware Medical Diagnostic Phrase Identification and Grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025).