

## Graphical Abstract

### **CX-Mind: A Pioneering Multimodal Large Language Model for Interleaved Reasoning in Chest X-ray via Curriculum-Guided Reinforcement Learning**

Wenjie Li, Yujie Zhang, Haoran Sun, Yueqi Li, Fanrui Zhang, Victoria Borja Clausich, Sade Mellin, Renhao Yang, Chenrun Wang, Zih-Shuo Wang, Shiyi Yao, Gen Li, Yidong Xu, Hanyu Wang, Yilin Huang, Angela Lin Wang, Chen Shi, Yin Zhang, Jianan Guo, Luqi Yang, Renxuan Li, Yang Xu, Jiawei Liu, Lei Liu, Bo Yan, Carlos Gutiérrez SanRomán, Lei Wang

## Highlights

### **CX-Mind: A Pioneering Multimodal Large Language Model for Interleaved Reasoning in Chest X-ray via Curriculum-Guided Reinforcement Learning**

Wenjie Li, Yujie Zhang, Haoran Sun, Yueqi Li, Fanrui Zhang, Victoria Borja Clausich, Sade Mellin, Renhao Yang, Chenrun Wang, Zih-Shuo Wang, Shiyi Yao, Gen Li, Yidong Xu, Hanyu Wang, Yilin Huang, Angela Lin Wang, Chen Shi, Yin Zhang, Jianan Guo, Luqi Yang, Renxuan Li, Yang Xu, Jiawei Liu, Lei Liu, Bo Yan, Carlos Gutiérrez SanRomán, Lei Wang

- Large chest X-ray dataset with over 2 million entries across 23 datasets.
- Novel training strategy boosts medical knowledge and reasoning in models.
- First interleaved reasoning approach for clear medical model interpretability.
- **CX-Mind** matches top medical reasoning models in extensive benchmark tests.
- Real-world clinical dataset validates **CX-Mind**'s utility with expert reviews.

# CX-Mind: A Pioneering Multimodal Large Language Model for Interleaved Reasoning in Chest X-ray via Curriculum-Guided Reinforcement Learning

Wenjie Li<sup>a,b,c,d,1</sup>, Yujie Zhang<sup>b,e,1</sup>, Haoran Sun<sup>d,f,1</sup>, Yueqi Li<sup>g</sup>, Fanrui Zhang<sup>h</sup>, Victoria Borja Clausich<sup>i</sup>, Sade Mellin<sup>j</sup>, Renhao Yang<sup>a,c</sup>, Chenrun Wang<sup>k</sup>, Zih-Shuo Wang<sup>l,m</sup>, Shiyi Yao<sup>c</sup>, Gen Li<sup>c</sup>, Yidong Xu<sup>c,1</sup>, Hanyu Wang<sup>c</sup>, Yilin Huang<sup>c</sup>, Angela Lin Wang<sup>c</sup>, Chen Shi<sup>c</sup>, Yin Zhang<sup>c</sup>, Jianan Guo<sup>c</sup>, Luqi Yang<sup>c</sup>, Renxuan Li<sup>c</sup>, Yang Xu<sup>c</sup>, Jiawei Liu<sup>h</sup>, Lei Liu<sup>f</sup>, Bo Yan<sup>e</sup>, Carlos Gutiérrez SanRomán<sup>o</sup>, Lei Wang<sup>a,c,\*</sup>

<sup>a</sup>College of Health Science and Technology, Shanghai Jiao Tong University School of Medicine, Shanghai, China

<sup>b</sup>Shanghai Innovation Institute, Shanghai, China

<sup>c</sup>Clinical Center for Sports Medicine, Department of Orthopaedics, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

<sup>d</sup>Shanghai Artificial Intelligence Laboratory, Shanghai, China

<sup>e</sup>Shanghai Key Laboratory of Intelligent Information Processing, School of Computer Science, Fudan University, Shanghai, China

<sup>f</sup>School of Basic Medical Sciences, Intelligent Medicine Institute, Fudan University, Shanghai, China

<sup>g</sup>Department of Hematology, The First Affiliated Hospital, College of Medicine, Zhejiang University, Hangzhou, China

<sup>h</sup>MoE Key Laboratory of Brain-Inspired Intelligent Perception and Cognition, University of Science and Technology of China, Hefei, China

<sup>i</sup>Department of Medicine, Faculty of Health Sciences, Universidad CEU Cardenal Herrera, Valencia, Spain

<sup>j</sup>Faculty of Medicine, University of Helsinki, Helsinki, Finland

<sup>k</sup>X-LANCE Lab, School of Computer Science, Shanghai Jiao Tong University, Shanghai, China

<sup>l</sup>Department of Hepatobiliary Surgery, National Cancer Center / National Clinical Research Center for Cancer / Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

<sup>m</sup>Department of Surgery, The Ohio State University Wexner Medical Center, The James Comprehensive Cancer Center, Columbus, United States

<sup>n</sup>Ningbo Institute of Technology, Beihang University, Ningbo, China

<sup>o</sup>Department of Pediatric Surgery (Servicio Cirugía Pediátrica), Hospital Universitario y Politécnico La Fe, Valencia, Spain

---

## Abstract

Chest X-ray (CXR) imaging is one of the most widely used diagnostic modalities in clinical practice, encompassing a broad spectrum of diagnostic tasks. Recent advancements have seen the extensive application of reasoning-based multimodal large language models (MLLMs) in medical imaging to enhance diagnostic efficiency and interpretability. However, existing multimodal models predominantly rely on "one-time" diagnostic approaches, lacking verifiable supervision of the reasoning process. This leads to challenges in multi-task CXR diagnosis, including protracted reasoning, sparse rewards, and difficulties in controlling hallucinations. To address these issues, we propose **CX-Mind**, the first generative model to achieve interleaved "think-answer" reasoning for CXR tasks, driven by curriculum-based reinforcement learning and verifiable process rewards (CuRL-VPR). Specifically, we constructed an instruction-tuning dataset, CX-Set, comprising 708,473 images and 2,619,148 samples, and generated 42,828 high-quality interleaved reasoning data points supervised by clinical reports. Optimization was

---

\*Corresponding author at College of Health Science and Technology, Shanghai Jiao Tong University School of Medicine, Shanghai, China . E-mail address: ray\_wangs@hotmail.com(L. Wang)

<sup>1</sup>Equal contributions.

conducted in two stages under the Group Relative Policy Optimization framework: initially stabilizing basic reasoning with closed-domain tasks, followed by transfer to open-domain diagnostics, incorporating rule-based conditional process rewards to bypass the need for pretrained reward models. Extensive experimental results demonstrate that **CX-Mind** significantly outperforms existing medical and general-domain MLLMs in visual understanding, text generation, and spatiotemporal alignment, achieving an average performance improvement of 25.1% over comparable CXR-specific models. On real-world clinical dataset (Rui-CXR), **CX-Mind** achieves a mean recall@1 across 14 diseases that substantially surpasses the second-best results, with multi-center expert evaluations further confirming its clinical utility across multiple dimensions. **CX-Mind** establishes a new paradigm for constructing interpretable, and high-performing medical MLLMs.

*Keywords:* Medical Reasoning; Multimodal Large Language Models; Chest X-Ray; Reinforcement Learning; Curriculum Learning

---

## 1. Introduction

In the past year alone, the global landscape of large language models (LLMs) has undergone a "reasoning revolution": OpenAI-o1 pioneered the introduction of extended chain-of-thought (CoT) pathways [1], while models such as DeepSeek-R1 and Gemini 2.5 Pro have continued to push the boundaries of reasoning capabilities [2, 3]. Empirical evidence has demonstrated that providing models with sufficient "reasoning space" enables them to exhibit near-expert-level proficiency in highly constrained tasks, such as mathematical proofs and structured programming. This momentum has rapidly transcended textual domains, extending into multimodal scenarios. When visual and linguistic signals are deeply integrated and mapped to a unified semantic space, CoT reasoning significantly amplifies model capabilities, consistently setting new benchmarks in cross-modal inference tasks [4].

In the medical domain, particularly in the high-frequency and task-diverse context of chest X-ray imaging, the development of user-friendly and interpretable AI assistants is imperative. Clinicians expect iterative feedback tracing evidence chains to conclusions, rather than opaque, one-shot answers. Only through reviewable and queryable intermediate conclusions can clinical personnel promptly identify and correct errors, ensuring safety and compliance [5, 6]. As shown in Figure 1(b), some studies have attempted to simulate "doctor-like reasoning" by leveraging large-scale chest X-ray report instruction fine-tuning or incorporating CoT prompts and reasoning process text in medical visual question answering (VQA) [7, 8]. As depicted in Figure 1(c), other research has employed reinforcement learning to enhance reasoning capabilities, designing rewards at the answer level or relying on additionally trained reward models [9, 10]. These efforts have significantly improved model performance in medical tasks, validating the value of CoT reasoning and reinforcement learning. However, existing medical reasoning multimodal large language models (MLLMs) predominantly adhere to a "one-shot judgment" paradigm, delivering final answers after protracted reasoning processes. This approach lacks verifiable process supervision, making it challenging for clinicians to intervene and correct errors in a timely manner [10, 11]. Furthermore, reliance on reward signals tied to final outcomes exacerbates the risk of hallucinations [12, 13]. Consequently, there is an urgent need in the chest X-ray domain for a novel paradigm that supports interleaved reasoning while mitigating hallucinations.

To address these challenges, we propose **CX-Mind**: a reasoning-focused multimodal large language model tailored for chest X-ray diagnostics, with "think-answer" interleaved reasoning as its fundamental interaction unit. This approach ensures interpretability and reviewability while significantly enhancing multi-task diagnostic performance. Specifically, we adopt

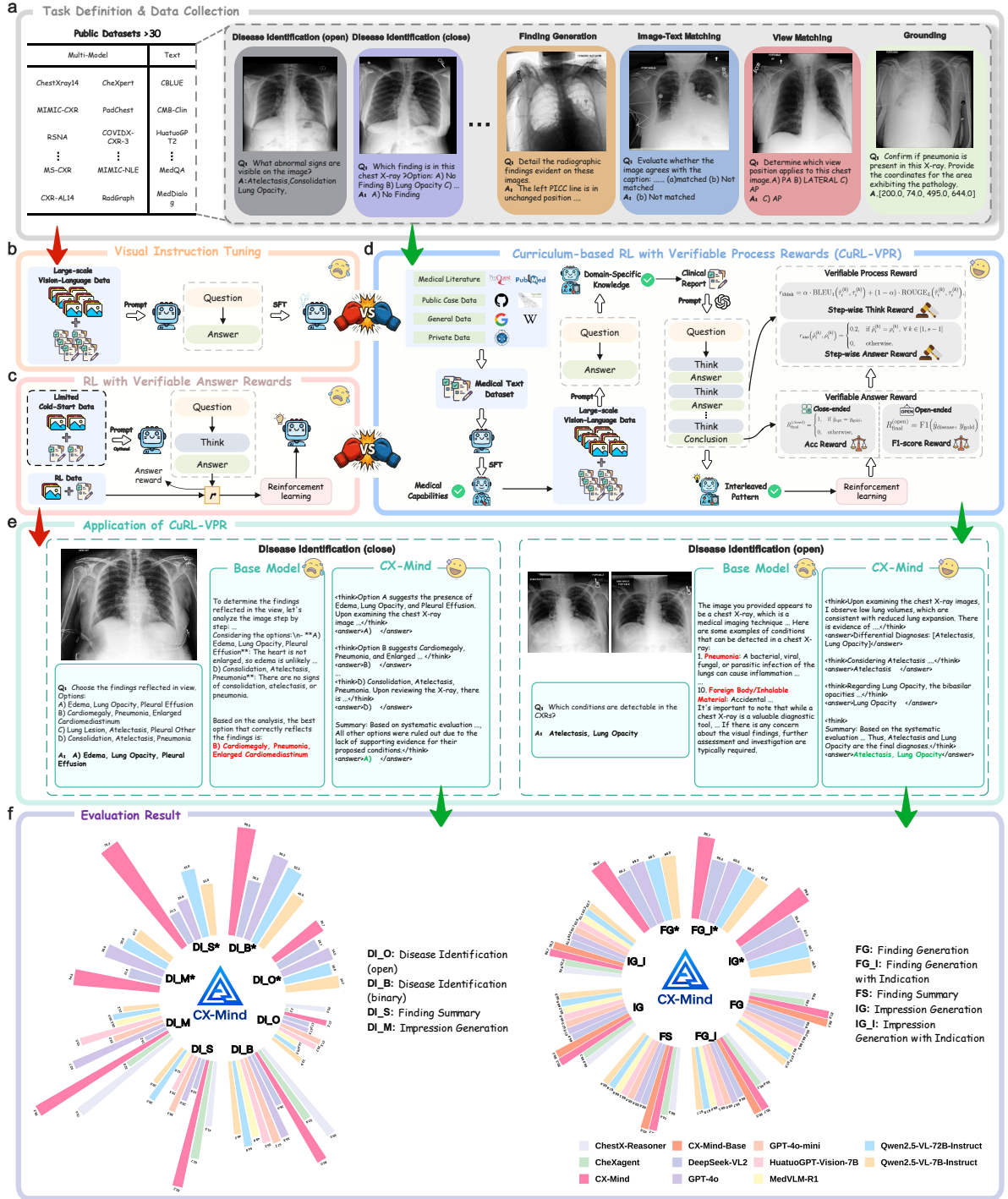


Figure 1: **CX-Mind: An MLLM for interleaved reasoning in Chest X-ray.** (a) Task Definition & Data Collection: A broad suite of chest X-ray interpretation tasks, with more than 30 single-model and multi-modal public datasets collected. (b) Visual Instruction Tuning: Fine-tuning with large-scale instruction data. (c) RL with Verifiable Answer Rewards: Reinforcement learning driven solely by outcome (answer) rewards. (d) CuRL-VPR framework: A curriculum-style RL fine-tuning paradigm that mines real reports to achieve process supervision. (e) Evaluation Result: Summary of CX-Mind versus numerous baselines, reporting Accuracy for visual understanding and BERTScore for text generation, and \* denotes out-of-domain testing.

curriculum-based reinforcement learning and verifiable process rewards as the core optimization strategy, with the overall workflow illustrated in Figure 1.

We first systematically outline three essential capabilities required for a chest X-ray reasoning model: Visual Understanding, Text Generation, and Spatiotemporal Alignment. As illustrated in Figure 1(a), inspired by [7], we decompose each capability into task subsets with clearly defined input-output structures. We collected over 30 publicly available datasets and, based on these, utilized DeepSeek-V3 [14] and GPT-4o [15] models to automatically construct a large-scale instruction dataset, CX-Set, comprising 708,473 images and 2,619,148 samples, along with 42,828 high-quality interleaved reasoning samples generated under supervision from real-world reports, providing robust data support for the interleaved reasoning paradigm.

Regarding training, as shown in Figure 1(d), we adopt a four-stage curriculum design. First, we leverage purely textual medical corpora to pre-train the language model component, enabling mastery of clinical terminology and reasoning patterns. Second, large-scale chest X-ray instruction fine-tuning injects vision-language knowledge and establishes robust semantic alignment for imaging data. Inspired by DeepSeek-R1 [3], reinforcement learning employs a two-stage training approach: the third stage uses a hybrid of answer-only and interleaved reasoning samples for cold-start supervision, allowing the model to learn the interleaved output format and providing a stable starting point for subsequent policy optimization. The fourth stage, conducted within the Group Relative Policy Optimization (GRPO) framework, similarly employs a curriculum-based approach for reinforcement learning. It begins with close-ended tasks to construct stable and verifiable reward signals, then transitions to open-ended diagnostics to achieve higher-level free reasoning capabilities [16]. Unlike traditional methods that rely solely on final answer rewards, our proposed verifiable process reward mechanism provides fine-grained feedback after each think-answer pair and employs conditional incentives, mitigating the credit assignment problem.

In the evaluation phase, we constructed a comprehensive benchmark covering open-source datasets such as MIMIC-CXR, CheXpert, and MS-CXR, as well as proprietary data from Rui-CXR. This benchmark includes over 30K VQA samples, assessing multiple tasks under each capability domain. As shown in Figure 1(f), experimental results demonstrate that **CX-Mind** significantly outperforms existing general-purpose and medical-specialized MLLMs across all three capability domains, achieving an average performance improvement of 25.1% compared to chest X-ray-specific models. More importantly, the interleaved output showcased in Figure 1(e) avoids the lengthy and difficult-to-review reasoning chains generated by traditional CoT approaches. Furthermore, we conducted comprehensive ablation studies and large-scale validation on the real-world clinical dataset Rui-CXR, accompanied by multidimensional evaluations from multiple clinical experts.

In summary, the main contributions of this paper can be outlined in six key points:

- We introduce a large-scale chest X-ray VQA dataset, CX-Set, encompassing 23 related datasets and 13 distinct tasks, with a total of over 2 million data entries. Additionally, we construct over 40K high-quality interleaved CoT data based on supervised patterns derived from real-world reports.
- We introduce a novel stage-wise training strategy guided by curriculum learning, which progressively injects chest X-ray-related medical knowledge and enhances the reasoning capabilities of MLLMs.
- We establish the first interleaved reasoning paradigm for multimodal medical applications, which enhances the interpretability of reasoning models and facilitates effective communication during human interaction.

- We design an effective rule-based process reward method to cultivate the predefined interleaved reasoning capabilities of MLLMs, with targeted training for both close-ended and open-ended questions.
- We demonstrate that **CX-Mind** achieves competitive performance compared to previous state-of-the-art medical reasoning MLLMs through extensive experiments across multiple benchmarks.
- We further conduct external validation using a collected real-world clinical dataset, Rui-CXR, and recruit three clinical experts to comprehensively evaluate **CX-Mind** across five distinct metrics to demonstrate its clinical utility.

## 2. Related Works

### 2.1. X-ray Foundation Model

Recent advancements in foundation models have revolutionized chest X-ray analysis, driving significant progress in image generation, disease diagnosis, and radiology report generation. These innovations leverage cutting-edge techniques such as domain-adapted pretraining and self-supervised learning to address persistent challenges like data scarcity and model generalizability. For instance, Chambon et al. [17] introduced a vision-language foundation model that generates highly realistic synthetic chest X-ray images by fine-tuning latent diffusion models on public datasets. Transitioning to disease diagnosis, Wang et al. [18] developed Ark+, an open-source model pretrained on diverse datasets with heterogeneous expert annotations. Ark+ achieves high diagnostic accuracy across multiple chest diseases while supporting privacy-preserving federated learning, making it a practical tool for clinical use. Complementing this, Zhang et al. [19] proposed EVA-X, which uses self-supervised learning to extract semantic and geometric features from unlabeled images. By reducing reliance on annotated data, EVA-X excels in over 11 detection tasks, advancing the field’s ability to address data scarcity. In the realm of radiology report generation, Chen et al. [20] introduced CheXagent, trained on the large-scale CheXinstruct dataset. CheXagent significantly reduces clinicians’ report drafting time, as validated by the CheXbench benchmark, enhancing clinical efficiency. Similarly, Lee et al. [21] compared M4CXR, a domain-specific MLLM, with ChatGPT-4o, showing that M4CXR outperforms in diagnostic accuracy and report generation due to its tailored adaptation to chest X-ray tasks. However, existing foundation models have not applied advanced techniques for enhancing reasoning capabilities to chest X-ray tasks. **CX-Mind**, through a multi-stage curriculum learning strategy, not only improves its ability to handle various fundamental clinical tasks but also enhances its reasoning capabilities for addressing complex problems.

### 2.2. Reasoning in Medical MLLMs

Advancements in LLMs have enhanced vision-language models, enabling them to address complex visual reasoning tasks that demand robust visual perception and advanced cognitive capabilities. Recent studies have optimized visual encoding strategies to improve the quality of visual tokens. Techniques such as prompt tuning, supervised fine-tuning (SFT), and reinforcement learning (RL) have proven critical in enhancing the reasoning capabilities of MLLMs. In medical domains, where interpretability is paramount, models like HuatuoGPT-o1 [22] and Baichuan-M1 [23] emphasize transparent reasoning paths to tackle medical challenges. However, integrating multimodal medical data for reasoning in MLLMs remains in its early stages. Well-crafted prompts simulating doctor-like reasoning processes show promise but are limited in applicability. Leveraging DeepSeek-R1’s success with RL, GRPO employs rule-based reward functions to improve adaptability across diverse medical tasks. For example, Med-R1 [9], utilizing GRPO, demonstrates superior performance across eight medical image modalities

and five question types, enhancing generalizability and reliability. Similarly, MedVLM-R1 [10] adopts an RL framework to promote interpretable reasoning paths without reliance on reference reasoning, mitigating overfitting issues associated with SFT. Despite these advancements, current models have yet to apply specialized medical reasoning to diagnostic tasks involving chest X-rays. Our proposed **CX-Mind** addresses this gap, enabling enhanced interaction with clinicians through interleaved reasoning processes.

### 2.3. RL for MLLMs Reasoning

RL has become a powerful approach for enhancing the reasoning capabilities of MLLMs, which integrate text and image data to perform tasks like visual question answering, image captioning, and clinical decision support. RL enables MLLMs to refine decision-making by optimizing reward functions, improving logical coherence and robustness over traditional SFT, which often struggles with overfitting and limited generalization. In general applications, RL techniques like GRPO guide MLLMs to focus on relevant visual and textual features, enhancing performance in complex reasoning tasks. In medical contexts, RL’s impact is particularly significant, as it supports interpretable and trustworthy reasoning across diverse data sources, such as medical images (e.g., X-rays, MRIs) and clinical notes. For instance, Lai et al. introduced Med-R1 [9], which leverages GRPO to achieve a 29.94% accuracy improvement over its base model (Qwen2-VL-2B) across eight medical imaging modalities, outperforming larger models and enhancing cross-task generalization by 32.06%. Similarly, Pan et al. proposed MedVLM-R1 [10], which uses RL to foster interpretable reasoning paths without reference reasoning, improving accuracy from 55.11% to 78.22% on MRI, CT, and X-ray benchmarks with minimal training data. Other works, such as Patho-R1 [24] and studies on radiology report generation, further demonstrate RL’s versatility in enhancing specialized medical reasoning. Although these methods integrate RL with medical tasks, they rely solely on the final answer as the basis for reward computation. In contrast, we propose a novel rule-based reward mechanism that encompasses the reasoning process, intermediate results, and the final answer. This approach is designed to train **CX-Mind**, thereby reducing hallucinations and enhancing reasoning accuracy.

## 3. Methodology

As described in Section 1, traditional CoT reasoning employs a "think-then-answer" approach, where the thinking component is often excessively verbose. We illustrate an example of the proposed interleaved diagnostic reasoning in Figure 2(c). Furthermore, existing multi-modal reasoning models only apply verifiable reinforcement learning to the final results, lacking supervision over the reasoning process. To address this, we propose curriculum-based reinforcement learning and verifiable process rewards (CuRL-VPR). Specifically, as shown in Figure 2, we introduce a fine-tuning paradigm for medical models that integrates SFT and open-ended reinforcement learning. This paradigm, grounded in a curriculum learning strategy, divides the training process into four stages. To ensure the verifiability and transferability of rewards, we exclusively employ rule-based multidimensional evaluators, mitigating the risk of reward hacking.

In the following, this section will elaborate on the curriculum learning strategy in a step-wise manner: (1) Section 3.2 introduces multiple metrics to filter high-quality textual medical data for enhancing foundational medical capabilities, (2) Section 3.3 constructs a large-scale instruction-following dataset, CX-Set, for knowledge injection, (3) Section 3.4 generates interleaved chain-of-thought data for cold-start training, (4) Section 3.5 implements curriculum-based reinforcement learning with process supervision.

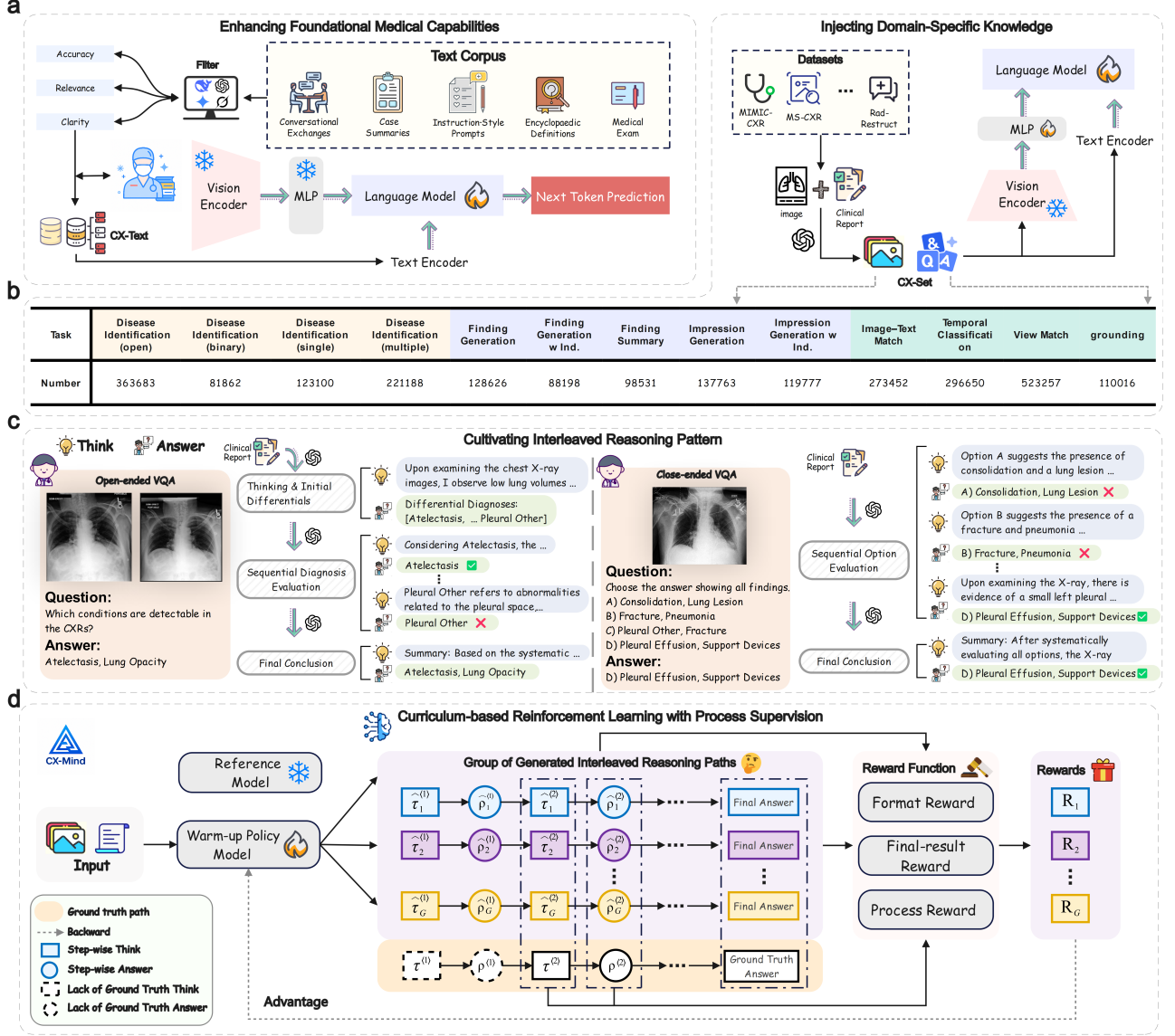


Figure 2: **CX-Mind training pipeline.** (a) Enhancing Foundational Medical Capabilities: We collect public text corpora and filter them with LLM-based scoring to obtain the CX-Text, then warm up the language-model component. (b) Injecting Domain-Specific Knowledge: We curate and construct over two million QA pairs to cultivate professional X-ray interpretation capability. (c) Cultivating the Interleaved Reasoning Pattern: Based on real medical reports, we prompt GPT-4o to build different forms of interleaved reasoning chains for close-ended and open-ended questions, thereby cold-starting the model. (d) RL with Process Supervision: The entire optimization is conducted with GRPO. The warm-up policy model  $\pi_\theta$  generates interleaved reasoning paths. Beyond outcome and format rewards, we further reward key reasoning steps.

### 3.1. Problem Definition

Unlike conventional approaches that rely on a vision-language model (VLM) to render a single-pass judgement, we introduce an interleaved diagnostic reasoning paradigm in which the model alternates between visual evidence parsing and linguistic inference, performing step-wise differential refinement to achieve both higher accuracy and stronger interpretability for chest X-ray analysis. Concretely, we construct a multimodal dataset  $D$ :

$$D = \left\{ (\mathbf{X}_i, \mathbf{C}_i, \mathbf{Q}_i, \mathbf{R}_i, \mathbf{S}_i, \mathbf{A}_i) \right\}_{i=1}^N, \quad (1)$$

where  $N$  is the number of studies,  $\mathbf{X}_i = \{x_i^{(1)}, \dots, x_i^{(K_i)}\}$  are the multi-view radiographs from the same examination,  $\mathbf{C}_i$  denotes the textual context,  $\mathbf{Q}_i$  contains both open-ended and close-ended questions,  $\mathbf{S}_i$  is the reference report, and  $\mathbf{R}_i = [(\tau_i^{(1)}, \rho_i^{(1)}), \dots, (\tau_i^{(T_i)}, \rho_i^{(T_i)})]$  is the ground-truth interleaved reasoning chain whose elements  $(\tau_i^{(s)}, \rho_i^{(s)})$  represent the  $s$ -th thought fragment and its accompanying answer snippet guided by  $\mathbf{S}_i$ , and  $\mathbf{A}_i$  records the definitive diagnosis and lesion coordinates.

We fine-tune an open-source multimodal large language model  $\pi_\theta$  so that, given only the images  $\mathbf{X}$ , context  $\mathbf{C}$ , and question  $\mathbf{Q}$ , it auto-regressively produces the interleaved sequence:

$$\mathcal{Y} = \underbrace{\tau^{(1)}}_{\text{think}} \oplus \underbrace{\rho^{(1)}}_{\text{answer}} \oplus \dots \oplus \tau^{(s)} \oplus \rho^{(s)} = \pi_\theta(\cdot \mid \mathbf{X}, \mathbf{C}, \mathbf{Q}), \quad (2)$$

where each pair  $(\tau^{(s)}, \rho^{(s)})$  is supervised by the corresponding ground-truth step in  $\mathbf{R}$ . The final answer fragment  $\rho^{(s)}$  must jointly realise lesion localisation, differential diagnosis, and report generation.

We partition the dataset as:

$$D = D_R \cup D_A, \quad D_R = D_{R_o} \cup D_{R_c}, \quad (3)$$

where  $D_R$  provides complete reasoning chains  $(\mathbf{R}, \mathbf{S}, \mathbf{A})$  and is further split into  $D_{R_o}$  (open-ended queries) and  $D_{R_c}$  (close-ended queries), whereas  $D_A$  contains answer-only samples without  $\mathbf{R}$ . The utilisation strategy of these subsets is detailed in subsequent section.

### 3.2. Enhancing Foundational Medical Capabilities

To equip the model with foundational medical vocabulary and clinical reasoning capabilities, we begin our curriculum with a concise, text-only SFT phase. Recent studies have indicated that such a targeted textual warm-up can significantly facilitate subsequent multimodal optimisation. For instance, HealthGPT reported improved stability and faster convergence following an initial SFT on clinical documents [25], while BiomedGPT observed notable performance improvements in medical vision-language tasks when the vision encoder was kept frozen during an early textual training stage [26]. Motivated by these findings, we restrict updates to the language layers in this stage, maintaining all visual parameters frozen in Figure 2(a). This ensures that the model internalises domain-specific knowledge before visual features are integrated.

The warm-up corpus comprises roughly 200K English question-answer (QA) pairs automatically translated from eight public resources. In terms of data types, it spans (i) conversational exchanges, (ii) structured case summaries, (iii) instruction-style prompts, (iv) encyclopaedic definitions, and (v) medical examination items. Concretely, we draw from CBLUE, CMB-Clin, HuatuoGPT2, CMExam, CMB-Exam, MedDialog and MedQA. This mixture exposes the model to a broad range of clinical entities, diagnostic cues, therapeutic options, and general medical reasoning patterns.

Each candidate QA pair is then evaluated by LLMs that serves as an automatic judge. We adopt a five-point scale on three enriched axes. **Accuracy** assesses whether the text faithfully

reflects medical facts, including symptoms, diagnoses, and treatment plans, and whether it is consistent with recognised clinical practice. **Relevance** measures how closely the content aligns with radiological diagnostic themes, penalising digressions into unrelated specialties. **Clarity** evaluates the use of standard medical terminology together with syntactic completeness and discourse coherence. Pairs scoring below three on any axis are discarded, and a random portion of the remaining data is independently reviewed by two senior physicians to confirm the automatic ratings.

### 3.3. Injecting Domain-Specific Knowledge

#### 3.3.1. Data Collection

Despite advancements in enhancing the model’s capability to address clinical medical problems, the complexity and diversity of medical tasks still prevent the model from achieving satisfactory performance in specialized domains. Theoretically, the data heterogeneity across different domains underscores the urgent need to incorporate domain-specific knowledge into the model. To lay the foundation for enhancing interleaved reasoning capabilities, we collected 23 publicly available datasets related to chest X-rays, including ChestXray14, CheXpert, MIMIC-CXR, PadChest, RSNA, COVIDX-CXR-3, CXR-LT, BRAX, NLM-TB, CandidPTX, BIMCV-COVID19, MS-CXR-T, VinDr-CXR, VinDr-PCXR, SIIM, MS-CXR, MIMIC-III, MIMIC-CXR-VQA, Rad-Restruct, MIMIC-NLE, RadGraph, MIMIC-Diff-VQA, and CXR-AL14. Each dataset typically contains one or more X-ray images from various perspectives, corresponding radiology reports, and common disease labels or visual localization coordinates. The diversity of data types is highly aligned with our objective of embedding X-ray-related knowledge into the model. Specifically, through comprehensive analysis of these extensive datasets, we identified three key capabilities that the model should possess: Visual Understanding, Text Generation, and Spatiotemporal Alignment.

#### 3.3.2. Construction of the CX-Set

To enhance the model’s performance across the three primary capabilities, we propose an instruction-fine-tuning dataset named CX-Set, tailored for chest X-rays. Specifically, we categorized different task types under each capability to comprehensively cover the knowledge domains in which the model should be trained. (i) **Visual Understanding:** Disease identification is the most common task in chest X-ray diagnostics, typically requiring differential diagnosis of 14 distinct diseases. Building on existing examination data, we constructed both open-ended and close-ended disease diagnosis questions, focusing on the interpretation and understanding of images. To diversify question types, close-ended questions are further classified into binary differentiation, single-disease diagnosis, and co-morbidity diagnosis. (ii) **Text Generation:** Writing radiology reports based on X-ray images is a time-consuming and repetitive task for clinicians. We aim to equip the model with the ability to generate specific sections of reports based on instructions, thereby alleviating the burden on clinicians. Specific tasks include findings generation, impression generation, and findings summary. (iii) **Spatiotemporal Alignment:** This encompasses additional core functionalities expected of an "X-ray expert", such as matching images with radiology reports, distinguishing images from different perspectives, assessing disease progression, and localizing diseases. Collectively, these tasks comprehensively address the capabilities required of an exceptional chest X-ray expert model.

As shown in Figure 2(b), We systematically constructed the CX-Set dataset based on the aforementioned tasks. To mitigate the impact of excessive noise on model performance, we first implemented an automated pipeline to clean the collected public datasets. This process involved removing data entries lacking images or radiology reports and filtering out cases with low-resolution images or reports that inadequately described disease conditions. Subsequently, to enhance data diversity, we designed 50 distinct question templates corresponding to the 13

previously defined task types. For task types where gold-standard answers could be directly extracted from the datasets, we adopted a “template-filling” approach for data construction [7]. Conversely, for tasks where answers could not be directly derived from existing data, we utilized DeepSeek-V3 [14] to generate standardized answers by summarizing radiology reports in alignment with specific questions. This automated data construction pipeline significantly reduced the labor-intensive manual annotation process. Leveraging this pipeline, we developed a large-scale chest X-ray instruction-fine-tuning dataset comprising 2,619,148 entries, covering 708,473 X-ray images. Each entry consists of three components: a task-specific question, one or more images, and the corresponding answer. To support enhanced interactive reasoning, we partitioned the entire dataset  $D$  into two subsets:  $D_A$  and  $D_R$ , which are used for injecting foundational X-ray knowledge and training subsequent reasoning capabilities, respectively. The objective of “knowledge injection” is to maximize the likelihood of generating correct answers based on the provided images and questions.

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(X,Q,A) \sim \mathcal{D}_A} \sum_{n=1}^N \log \pi_{\theta}(a_n \mid X, Q, a_{<n}), \quad (4)$$

Where  $X$ ,  $Q$ , and  $A$  represent the image, question, and corresponding answer, respectively.  $\pi_{\theta}$  denotes the policy model under training,  $n$  represents the index of the token  $a$  decoded by the model, and  $N$  corresponds to the length of  $A$ .

### 3.4. Cultivating Interleaved Reasoning Pattern

#### 3.4.1. Definition of Interleaved Reasoning

Interleaved reasoning in medical MLLMs entails addressing complex, multi-step medical queries by resolving sub-tasks step-by-step. This approach decomposes a query into a series of intermediate reasoning steps, each producing a distinct, user-facing “sub-answer” that represents a confident, self-contained conclusion or milestone in the reasoning process. For example, in the diagnosis of a chest X-ray, a sub-answer might identify a specific abnormality, such as a lung nodule, as a resolved step, guiding subsequent reasoning toward differential diagnoses or treatment recommendations.

A critical distinction in this framework is between thinking and answering. Thinking refers to the internal, private reasoning processes—such as hypothesis generation or cross-modal data integration that are not directly accessible or immediately useful to the user [27]. In contrast, answering involves generating public, finalized outputs that advance the user’s understanding or support clinical decision-making. These outputs, presented as sub-answers, are conclusive at their respective stages and collectively contribute to a comprehensive response. For instance, in a multi-hop medical query requiring both image analysis and clinical correlation, a sub-answer might confirm a visual finding (e.g., “The X-ray reveals pulmonary edema”) before proceeding to evidence-based differential diagnoses supported by image analysis. By interleaving reasoning across modalities and delivering clear, incremental sub-answers, medical MLLMs enhance transparency, interpretability, and collaboration with clinicians.

Specifically, we designed distinct interleaved reasoning modes based on the type of question in Figure 3(c). For close-ended questions, **CX-Mind** systematically evaluates each option to ensure that it is either retained or excluded based on evidence. Subsequently, it summarizes the reasoning process and intermediate results to derive the final conclusion. Conversely, when addressing open-ended questions, **CX-Mind** first identifies potential diseases based on a preliminary analysis of the image. It then evaluates each of these diseases through an evidence-based approach, ultimately providing a diagnostic conclusion. Formally, the interleaved reasoning mode is defined as follows:

$$\mathcal{Y}_{\text{close}} = \underbrace{\tau^{(1)} \oplus \rho^{(1)}}_{\text{Option A}} \oplus \underbrace{\tau^{(2)} \oplus \rho^{(2)}}_{\text{Option B}} \oplus \dots \oplus \underbrace{\tau^{(s)} \oplus \rho^{(s)}}_{\text{Final Answer}} \quad (4)$$

$$\mathcal{Y}_{open} = \underbrace{\tau^{(1)} \oplus \rho^{(1)}}_{\text{Possible List}} \oplus \underbrace{\tau^{(2)} \oplus \rho^{(2)}}_{\text{Disease 1}} \oplus \dots \oplus \underbrace{\tau^{(s)} \oplus \rho^{(s)}}_{\text{Final Answer}} \quad (5)$$

where  $\mathcal{Y}$  represents the interleaved reasoning process, and  $\tau$  and  $\rho$  denote the reasoning and answer segments, respectively, corresponding to Equation 2.

### 3.4.2. Cold Start for Interleaved Pattern

To enhance the interleaved reasoning capabilities of **CX-Mind**, we initially employed a cold-start phase to enable the model to familiarize itself with and master the interleaved reasoning mode. This approach lays the groundwork for subsequent RL to further elevate reasoning performance. Specifically, we constructed 42,828 high-quality interleaved CoT data entries, guided by tailored prompts and supplemented by authentic radiology reports as supervisory signals. These data encompass complex tasks, including various types of disease identification and disease progression. The CoT data construction pipeline is outlined as follows:

**(i) Report Screening:** As authentic radiological findings in reports provide critical support for addressing complex tasks, we extracted information from the "Findings" section. However, the absence of a standardized format in the collected clinical reports posed challenges to extracting specific sections. To streamline this process, we targeted only reports containing the strings "FINDINGS:" and "IMPRESSION:", extracting the content between these markers to serve as supervisory signals for subsequent steps.

**(ii) Token Statistics:** More detailed radiological findings offer richer supportive evidence to guide the generation of high-quality CoT data. Accordingly, we utilized the tokenizer from the Qwen3 series to perform token counting on the radiology reports. Data entries with a "Findings" section exceeding 120 tokens were selected as the raw data for CoT construction.

**(iii) Category Balancing:** Statistical analysis revealed an imbalanced label distribution in the collected data. For instance, co-morbidity diagnosis tasks included a disproportionate number of entries labeled "No Finding." To enhance the robustness of reasoning capabilities, we employed an automated approach to balance the label distribution across different task types.

**(iv) Interleaved Reasoning Generation:** Leveraging the advanced language and vision capabilities of ChatGPT-4o, we generated interleaved reasoning processes for various task types. Notably, unlike other CoT construction methods, we utilized the authentic radiological findings extracted earlier as supervisory signals to minimize hallucinations. Specifically, we applied distinct interleaved reasoning logics tailored to different task types to align with practical scenarios. The reasoning logic for open-ended questions is illustrated in Figure 3:

- - Open-Ended: Thinking & Initial Differentials  $\longrightarrow$  Sequential Diagnosis Evaluation  $\longrightarrow$  Final Conclusion
- - Close-Ended: Sequential Option Evaluation  $\longrightarrow$  Final Conclusion
- - Binary: Analysis and Conclusion

During the cold-start phase, we employed SFT to enable **CX-Mind** to adapt to the interleaved reasoning mode by utilizing a hybrid dataset composed of  $D_A$  and a subset of  $D_R$ . The dataset  $D_R$  was randomly divided into  $D_{R1}$  and  $D_{R2}$ , each comprising half of the data. The former was utilized for cold-start training, with the objective function defined as follows:

$$\mathcal{L}_{\text{Cold-Start}} = -\mathbb{E}_{(X, Q, C, R) \sim \mathcal{D}_{R1}} \sum_{n=1}^N \log \pi_{\theta}(r_n \mid X, Q, C, r_{<n}), \quad (7)$$

where  $C$  and  $R$  represent the instruction output in a specific format and the complete interactive reasoning process, respectively.

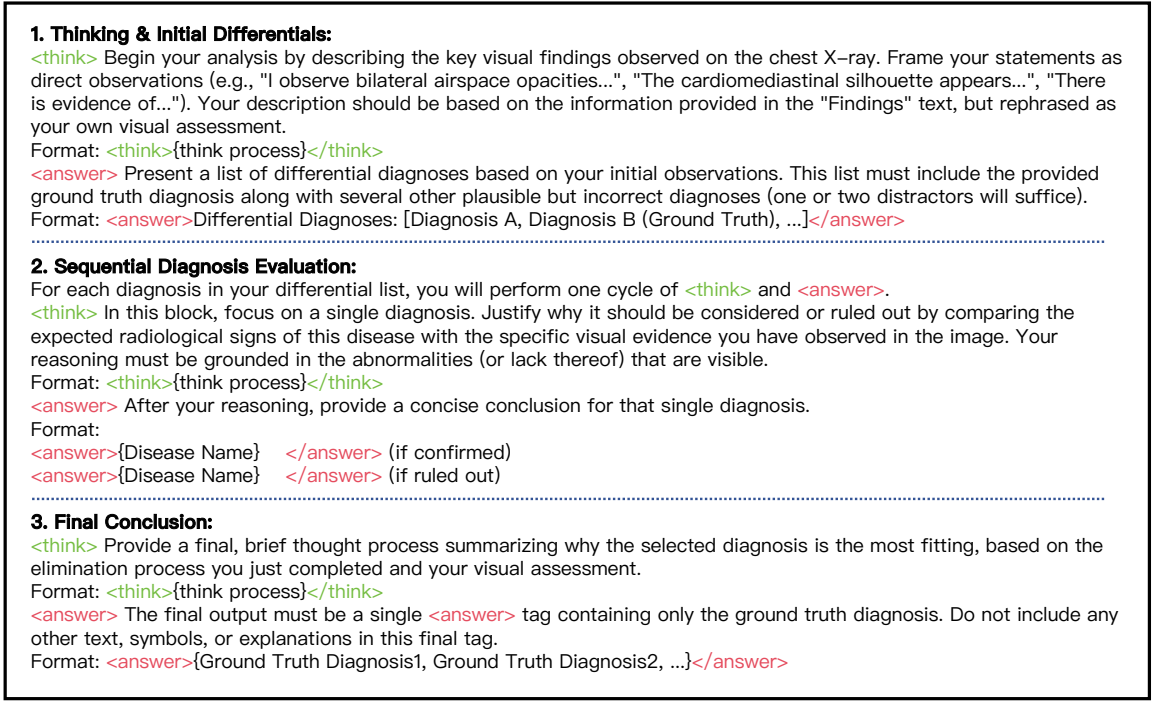


Figure 3: Illustration of the interleaved reasoning process required for open-ended questions.

### 3.5. Curriculum-based Reinforcement Learning with Process Supervision

This section details our curriculum-based reinforcement learning pipeline. We adopt a GRPO backbone and extend it with step-wise process rewards so that the policy is updated not only on the final answer but after every think-answer pair. Rewards are shaped to capture local consistency, cross-turn coherence and medical safety within the interleaved stream. Training proceeds in two phases: the policy is first stabilised on close-ended questions and then unlocked for open-ended queries, enabling a smooth transition from constrained to free-form reasoning.

#### 3.5.1. GRPO-Based Optimization Methodology

With the release of Deepseek-R1, GRPO has been shown to markedly enhance the reasoning capabilities of LLMs. Recent studies further demonstrate that the GRPO algorithm can improve the cross-modal reasoning ability of medical vision-language models. As a variant of Proximal Policy Optimization (PPO), GRPO computes the generalized advantage estimate using group-relative rewards. Unlike methods that rely on pretrained reward models, GRPO evaluates outputs through a verifiable rule-based procedure. Specifically, for each sample  $(\mathbf{X}_i, \mathbf{C}_i, \mathbf{Q}_i, \mathbf{R}_i, \mathbf{S}_i, \mathbf{A}_i)$  in  $D$ , the model generates  $G$  candidate responses  $\{o_i\}_{i=1}^G$  under the current policy  $\pi_\theta$ . Each response receives a corresponding reward  $r_i$ , and the group of rewards is normalized to compute the advantage:

$$A_i = \frac{r_i - \text{mean}(\{r_j\}_{j=1}^G)}{\text{std}(\{r_j\}_{j=1}^G)}. \quad (6)$$

The GRPO objective can be written as:

$$J(\theta) = \mathbb{E}_{q \sim P(Q), \{o_i\}} \left[ \frac{1}{G} \sum_{i=1}^G \min(R_i A_i, \text{clip}(R_i, 1 - \epsilon, 1 + \epsilon) A_i) - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \right], \quad (7)$$

where  $R_i = \frac{\pi_\theta(o_i | \mathbf{X}, \mathbf{C}, \mathbf{Q})}{\pi_{\theta_{\text{old}}}(o_i | \mathbf{X}, \mathbf{C}, \mathbf{Q})}$ , and its corresponding advantage  $A_i$ . The clipping mechanism limits extreme updates by constraining  $R_i$  to the interval  $[1 - \epsilon, 1 + \epsilon]$ , while the KL-divergence term weighted by  $\beta$  penalizes large deviations from the reference policy  $\pi_{\text{ref}}$ .

### 3.5.2. Rule-based Rewards Policy for Interleaved Process

Because GRPO lacks inherent supervision on intermediate steps and pre-trained reward models are prone to reward hacking, we define explicit rule-based signals for both close-ended and open-ended questions in Figure 2(d). To ensure correct parsing of the interleaved response, we prompt the model during training with the instruction template shown in Figure 4.

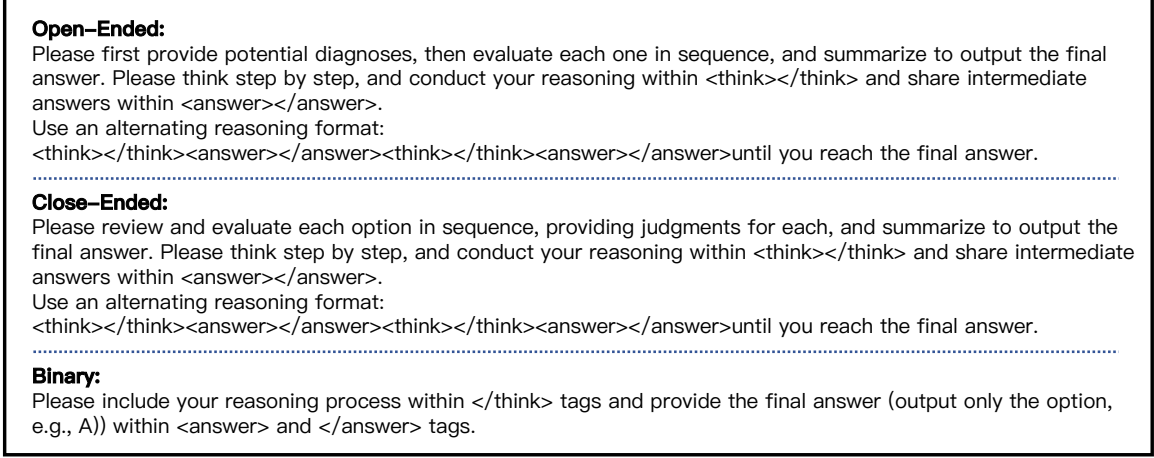


Figure 4: Prompts following specific thinking-answer formats for different types of questions.

- **Format Reward  $R_{\text{format}}$ :** A trajectory earns 1.0 if every `<think>`–`</think>` fragment is immediately followed by a matching `<answer>`–`</answer>` fragment and the sequence ends with an `<answer>`, otherwise 0.0.
- **Final-result Reward  $R_{\text{final}}$ :** The final accuracy reward evaluates the correctness of the last `<answer>` segment. For close-ended questions we use a binary reward:

$$R_{\text{final}}^{(\text{closed})} = \begin{cases} 1, & \text{if } \hat{y}_{\text{opt}} = y_{\text{gold}}, \\ 0, & \text{otherwise,} \end{cases} \quad (8)$$

where  $\hat{y}_{\text{opt}}$  is the predicted choice and  $y_{\text{gold}}$  is the ground-truth label for a multiple-choice question. For open-ended diagnosis, a single choice is often insufficient because multiple conditions may co-occur. We therefore employ the micro-F1 score, which balances precision and recall over the entire disease set:

$$R_{\text{final}}^{(\text{open})} = \text{F1}(\hat{y}_{\text{disease}}, y_{\text{gold}}), \quad (9)$$

where  $\hat{y}_{\text{disease}}$  and  $y_{\text{gold}}$  denote the predicted and reference disease sets, respectively.

- **Process Reward  $R_{\text{proc}}$ :** To harness real reports for supervising multi-step reasoning and curb hallucinations, we first verify the final answer and only then allocate rewards to intermediate steps; blindly rewarding the whole chain often fails because of the credit-assignment problem. Instead, our mechanism first verifies the answer and then assigns an additional bonus to logically consistent and well-structured reasoning traces. Inspired by [27], we introduce a conditional, interleaved reward scheme that encourages correct reasoning. The method is entirely rule-based and multi-step, requiring no trained reward model, which makes it both verifiable and efficient.

Concretely, a process reward is issued for close-ended questions only when three priors hold: (i) the output format is correct, (ii) the final answer is correct, and (iii)

the mini-batch accuracy exceeds the exponential moving average (EMA) of all previous batches. This design ensures that intermediate supervision is applied only after the model shows meaningful learning progress. For open-ended questions, conditions (ii) and (iii) are combined by replacing accuracy with an F1-score. Formally, let  $\text{Acc}_b$  be the accuracy of mini-batch  $b$  and  $\text{EMA}_{b-1}$  the EMA from preceding batches. Only samples satisfying all conditions receive the additional rewards  $r_{\text{think}}$  and  $r_{\text{ans}}$  for their intermediate reasoning steps and corresponding answers. The conditional process reward is:

$$R_{\text{proc}}(\{\hat{\tau}_i^{(k)}, \hat{\rho}_i^{(k)}\}_{k=1}^{s-1}, \{\tau_i^{(k)}, \rho_i^{(k)}\}_{k=1}^{s-1}, \hat{y}, y) \\ = \mathbb{I}(C) \sum_{k=1}^{s-1} r_{\text{think}}(\hat{\tau}_i^{(k)}, \tau_i^{(k)}) + r_{\text{ans}}\{\hat{\rho}_i^{(k)}, \rho_i^{(k)}\}_{k=1}^{s-1}, \quad (10)$$

$$C = \text{Check}(\{\hat{\tau}_i^{(k)}, \hat{\rho}_i^{(k)}\}_{k=1}^s) \wedge \text{Judge}(\hat{y}) \wedge (\text{Acc}_b > \text{EMA}_{b-1}), \quad (11)$$

where  $\mathbb{I}(\cdot)$  denotes the indicator function,  $\hat{y}$  is the predicted-answer segment, and  $\text{Check}(\cdot)$  and  $\text{Judge}(\cdot)$  verify the interleaved format and the correctness of the final answer, respectively. For close-ended questions,  $\text{EMA}_{b-1}$  is iteratively updated with each batch’s accuracy. For open-ended questions, the same criterion applies with accuracy replaced by the batch-level F1-score.

For the thinking reward  $r_{\text{think}}$ , the ground-truth reasoning text is drawn verbatim from authoritative medical reports. We therefore devise a text-similarity-based reward function: each model-generated reasoning part is aligned with its ground-truth counterpart, after which BLEU-1 and ROUGE-L are computed to quantify lexical overlap. Their weighted average constitutes the reward:

$$r_{\text{think}} = \alpha \cdot \text{BLEU}_1(\hat{\tau}_i^{(k)}, \tau_i^{(k)}) + (1 - \alpha) \cdot \text{ROUGE}_L(\hat{\tau}_i^{(k)}, \tau_i^{(k)}), \quad (12)$$

where  $\alpha \in [0, 1]$  balances the two metrics.

For  $r_{\text{ans}}$ , we adopt an All-or-None self-consistent reward: a fixed coefficient  $\gamma$  is granted only when every intermediate conclusion fragment is semantically identical to the reference answer, and no extra score is given if any discrepancy exists.

$$r_{\text{ans}}(\hat{\rho}_i^{(k)}, \rho_i^{(k)}) = \begin{cases} 0.2, & \text{if } \hat{\rho}_i^{(k)} = \rho_i^{(k)}, \forall k \in [1, s-1], \\ 0, & \text{otherwise.} \end{cases} \quad (13)$$

This minimalist rule constrains answer-level accuracy while preventing misleading incentives arising from partial matches.

- **Total Reward  $R$ :** The overall per-trajectory reward can be formulated as the weighted sum:

$$R = \lambda R_{\text{format}} + (1 - \lambda) R_{\text{final}} + R_{\text{proc}}, \quad (14)$$

Specifically,  $\lambda$  modulates the relative contributions of the format reward and the final reward, where  $R_{\text{final}}$  is instantiated as  $R_{\text{final}}^{(\text{closed})}$  for close-ended questions and as  $R_{\text{final}}^{(\text{open})}$  for open-ended questions.

### 3.5.3. Training Strategy

In contrast to fields such as mathematics or formal logic, where the answer space is strictly defined, the reward landscape of open-ended medical diagnosis is extremely sparse, which greatly reduces the efficiency of RL updates. Drawing on curriculum learning [28] and further inspired by the approach in [16], we first train the agent on simpler close-ended tasks, such as binary judgements and multiple-choice questions, to establish fundamental reasoning skills. We then refine the policy on open-ended diagnostic tasks. The complete reinforcement-learning pipeline is summarised in Algorithm 1.

## 4. Experiments

### 4.1. Dataset

To demonstrate the broad generalization capabilities of **CX-Mind**, we evaluated its multimodal chest X-ray abilities and foundational clinical language proficiency. Specifically, for the three designated capability categories—visual understanding, text generation, and spatiotemporal alignment—we utilized tasks constructed from MIMIC-CXR [29] and CheXpert [30] as training and in-domain evaluation datasets. For each task category, 2,000 samples were randomly selected as the test set, with the remainder used for training. Correspondingly, we established an out-of-domain test set using OpenI [31], following the same data construction methodology (500 samples per task category), to validate the model’s robust generalization. For the evaluation of language proficiency, we selected datasets [32] from five categories to comprehensively assess **CX-Mind**, encompassing medical language understanding (CHIP-CDN, CMeEE), medical language generation (IMCS-V2-MRG), complex medical reasoning (DDx-basic), medical safety and ethics (MedSafety), and medical knowledge question-answering (MedHG, Med-Exam). These datasets were randomly split, with 90% of the data allocated for training and the remaining 10% for testing.

### 4.2. Evaluation Metrics

For the evaluation metrics, we employed distinct measures tailored to the task types. For close-ended questions, accuracy (Acc) was used as the primary metric. For open-ended questions, disease identification tasks were assessed using the Jaccard index and accuracy, with a threshold of Jaccard  $> 0.5$ . For report generation tasks, we adopted standard natural language generation metrics, including BLEU, ROUGE-1, ROUGE-2, and ROUGE-L. Additionally, BERTScore was utilized to evaluate semantic similarity. For disease localization tasks, we used Intersection over Union (IoU) and accuracy, with a threshold of IoU  $> 0.5$ .

### 4.3. Implementation Details

The entire training pipeline, encompassing SFT and RL, was conducted using eight 80GB H100 GPUs. During the SFT phase, we adopted QwenVL2.5-VL-7B-Instruct as the base model for **CX-Mind**, employing LoRA for fine-tuning, with AdamW as the optimizer, a learning rate of  $1e-4$  following a cosine decay schedule, a batch size of 16, and bfloat16 mixed precision. For the RL phase, we employed the EasyR1 framework to achieve efficient VLM training, leveraging the HybridEngine from verL. The training utilized the model after SFT for full-parameter fine-tuning, with FlashAttention-2 implemented to enhance efficiency. The rollout batch size was set to 64, with a maximum input pixel limit of 4,194,304. The learning rate was configured at  $1e-6$ , with a weight decay of 0.01 and a KL penalty coefficient  $\beta$  of 0.01, with  $\lambda = 0.2$  and  $\alpha = 0.3$ . For each sample, 10 completions were generated using vLLM with a sampling temperature of 1.0. RL training was conducted for 1 epoch on close-ended questions and 2 epochs on open-ended questions, totaling approximately 12 hours.

---

**Algorithm 1** Curriculum GRPO Training with Rule-Based Process Rewards

---

**Require:** **Datasets:** close-ended  $\mathcal{D}_{\text{closed}}$  and open-ended  $\mathcal{D}_{\text{open}}$ ; **Policies:** initial  $\pi_\theta$  and reference policy  $\pi_{\text{ref}}$ ; **GRPO hyper-parameters:** group size  $G$ , clipping window  $\epsilon$ , KL weight  $\beta$ ; **Reward weights:** format-final trade-off  $\lambda$ , thinking-similarity weight  $\alpha$ , process bonus  $\gamma$ ; **Training steps:**  $N_c$  (closed phase) and  $N_o$  (open phase).

*// Method*

```
1: function TRAINPHASE( $\mathcal{D}$ ,  $N$ , CLOSED)
2:   Initialise EMA  $\leftarrow 0$ 
3:   for  $t = 1$  to  $N$  do ▷ Iterate Mini-batches
4:     Sample batch  $B \subset \mathcal{D}$ ;  $L_{\text{PG}} \leftarrow 0$ 
5:     for all query  $q \in B$  do
6:       Generate  $G$  trajectories  $\{o_j\}_{j=1}^G \sim \pi_\theta(o \mid q)$ 
7:       for  $j = 1$  to  $G$  do ▷ Rule-based Reward
8:          $R_{\text{format}} \leftarrow \mathbb{I}(\text{interleaved format correct})$ 
9:         if CLOSED then ▷ Final-result Reward
10:           $R_{\text{final}} \leftarrow \mathbb{I}(\hat{y}_{\text{opt}} = y_{\text{gold}})$ 
11:        else
12:           $R_{\text{final}} \leftarrow \text{F1}(\hat{y}_{\text{disease}}, y_{\text{gold}})$ 
13:        end if
14:         $C \leftarrow (R_{\text{format}} = 1) \wedge (R_{\text{final}} > 0) \wedge (\text{BatchMetric} > \text{EMA})$ 
15:        if  $C$  then ▷ Process Reward
16:           $R_{\text{proc}} \leftarrow \sum_{k=1}^{s-1} r_{\text{think}}^{(k)} + r_{\text{ans}}$ 
17:        else
18:           $R_{\text{proc}} \leftarrow 0$ 
19:        end if
20:         $r_j \leftarrow \lambda R_{\text{format}} + (1 - \lambda) R_{\text{final}} + R_{\text{proc}}$ 
21:      end for
22:       $\mu, \sigma \leftarrow \text{mean/std}(\{r_j\}); A_j = (r_j - \mu) / \sigma$ 
23:       $L_{\text{PG}} += \frac{1}{G} \sum_{j=1}^G \min(R_j A_j, \text{clip}(R_j, 1 - \epsilon, 1 + \epsilon) A_j)$ 
24:    end for
25:     $\theta \leftarrow \theta + \eta \nabla_\theta (L_{\text{PG}} / |B| - \beta D_{\text{KL}}(\pi_\theta \parallel \pi_{\text{ref}}))$ 
26:    Update EMA with current batch accuracy (or F1)
27:  end for
28: end function
// Training Pipeline
29: Phase 1: close-ended RL
30: TRAINPHASE( $\mathcal{D}_{\text{closed}}$ ,  $N_c$ , CLOSED = True)
31: Phase 2: Open-ended RL
32: TRAINPHASE( $\mathcal{D}_{\text{open}}$ ,  $N_o$ , CLOSED = False)
```

---

#### 4.4. Baselines

To evaluate the performance of **CX-Mind**, we compare it against the following three categories of baseline models: **(i) General MLLMs:** This category includes high-performance vision-language models designed for general-purpose tasks. We selected both closed-source models, such as GPT-4o [15] and GPT-4o-mini [15], known for their robust performance across diverse benchmarks including visual question answering and reasoning, and open-source models, such as Qwen2.5-VL-7B-Instruct [33], Qwen2.5-VL-72B-Instruct [33] and DeepSeek-VL2 [34], which demonstrate advanced capabilities in image and video understanding, document parsing, and visual agent tasks. **(ii) Medical MLLMs:** This category comprises vision-language models pretrained or fine-tuned on medical corpora, tailored for medical imaging and clinical tasks. We include HuatuoGPT-Vision-7B [22], which leverages medical-specific data for enhanced clinical understanding, CheXagent [20], designed for chest X-ray analysis. **(iii) Medical Reasoning MLLMs:** This category includes models specifically fine-tuned for advanced medical reasoning tasks, often incorporating techniques such as GRPO. We selected MedVLM-R1 [10], optimized for medical reasoning, and ChestX-Reasoner [35], which focuses on reasoning tasks specific to chest X-ray diagnostics. These baseline models provide a comprehensive comparison framework, encompassing general-purpose multimodal capabilities, medical-specific pretraining, and advanced medical reasoning, to rigorously evaluate the performance and generalization of **CX-Mind** across diverse chest X-ray tasks.

#### 4.5. Main Results

##### 4.5.1. Performance on Visual Understanding

We first evaluated the performance of **CX-Mind** in visual understanding capabilities, comparing it with other SOTA models. Specifically, close-ended and open-ended disease identification tasks were employed to assess the model’s ability to interpret X-ray images and detect abnormalities. As disease identification is a fundamental skill expected of radiologists, we categorized the tasks into binary (yes/no) classification, single-disease identification, co-morbidity identification, and free-form responses without predefined options, aiming to enhance the robustness of the evaluation results. Furthermore, we conducted evaluations on both in-domain (MIMIC-CXR, CheXpert) and out-of-domain (OpenI) test sets.

As shown in Figure 5, the results demonstrate that **CX-Mind** exhibits superior performance across various disease identification tasks, outperforming a range of large-scale closed-source and open-source models, such as GPT-4o and Qwen2.5-VL-72B-Instruct, as well as models fine-tuned on specific X-ray corpora, such as ChestX-Reasoner and CheXagent. This performance advantage remains consistent in out-of-domain test sets, underscoring **CX-Mind**’s robust visual understanding capabilities. Specifically, for binary classification tasks, **CX-Mind** outperformed the best baseline model, ChestX-Reasoner, by 15% (82% vs. 67%) and 6% (78% vs. 72%) in in-domain evaluations. In out-of-domain testing, **CX-Mind** achieved a 12.2% performance improvement over GPT-4o. Due to the simplicity of binary tasks, **CX-Mind**’s advantage is less pronounced in this context. However, for single-disease and co-morbidity identification tasks, the performance gap widens significantly. Compared to models fine-tuned on chest X-ray datasets, **CX-Mind** achieved an average performance improvement of 19.5% and 21% over CheXagent and ChestX-Reasoner, respectively, in single-disease identification, and 63.5% and 21.2% in co-morbidity diagnosis. This highlights **CX-Mind**’s substantial advantage in complex tasks such as co-morbidity detection, an area often overlooked by comparable models. For open-ended tasks, we required all models to output results in a list format, with ChatGPT-4o used to extract answers to account for non-standardized outputs. The results indicate that most models experience a significant performance drop when no options are provided, with the exception of **CX-Mind** and ChestX-Reasoner. Compared to the latter, **CX-Mind** achieved

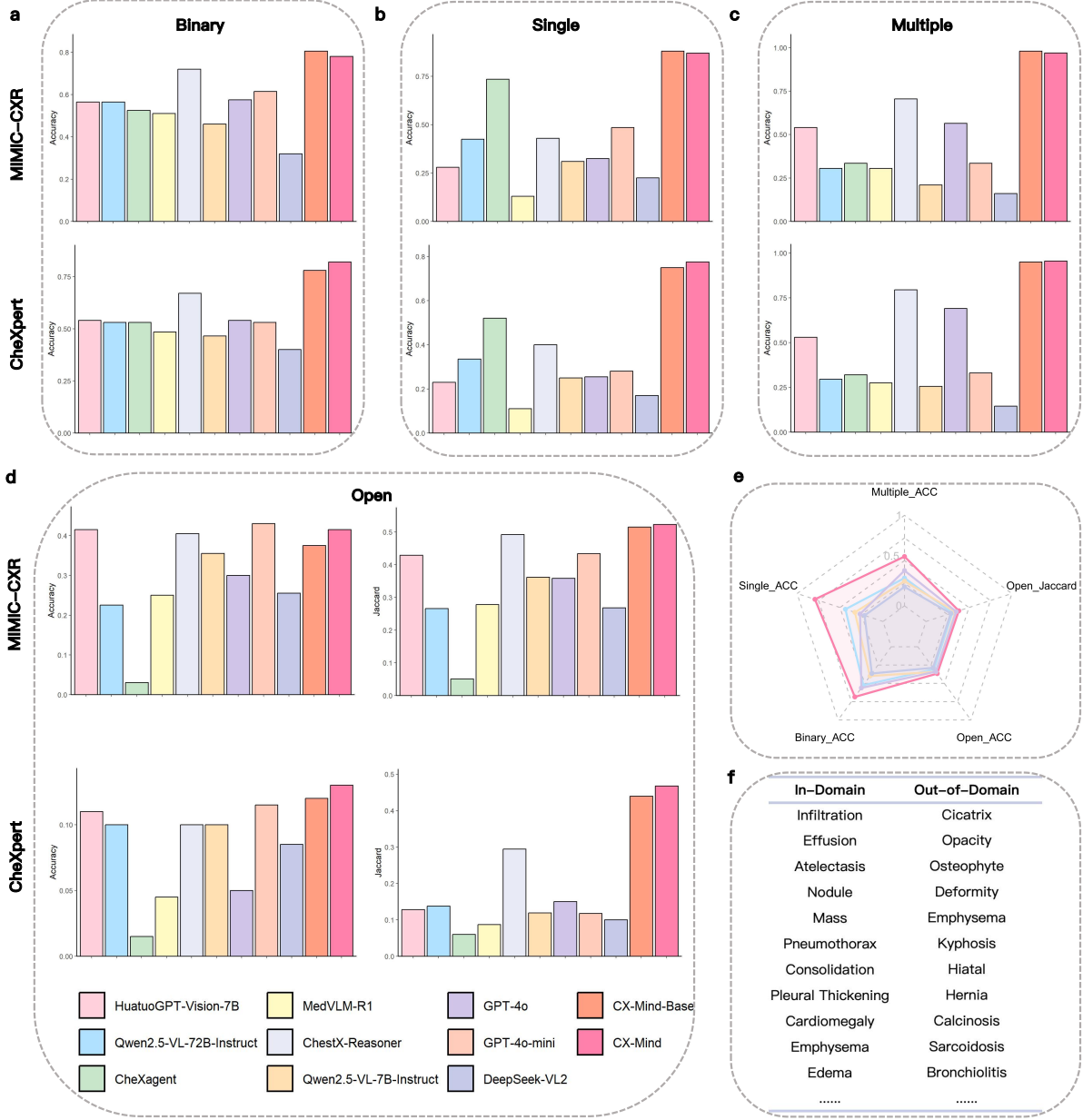


Figure 5: Comparative study of CX-Mind and various baseline models in visual understanding capabilities. (a) Performance comparison of different models on a binary disease classification task, where the in-domain test set is primarily derived from authentic reports in MIMIC-CXR and CheXpert. (b) Performance comparison of different models on a single-disease identification task. (c) Performance comparison of different models on a co-morbidity identification task. (d) Performance comparison of different models on an open-ended disease identification task, evaluated using precision and Jaccard score. (e) Performance comparison of CX-Mind against mainstream models on an out-of-domain test set (OpenI). (f) Differences in disease categories covered by in-domain and out-of-domain test sets.

performance improvements of 3.1% and 17.2% on the MIMIC-CXR and CheXpert test sets, respectively.

#### 4.5.2. Performance on Text Generation

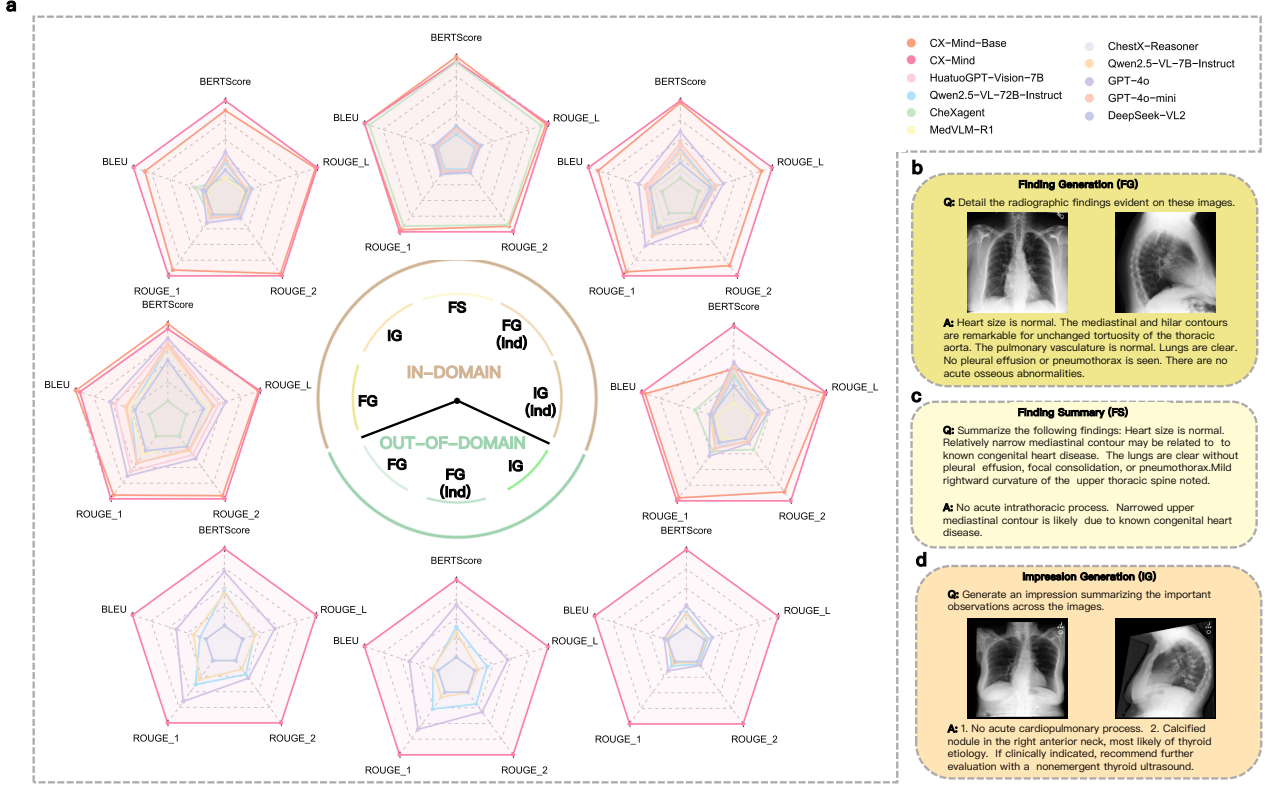


Figure 6: Comparative study of CX-Mind and various baseline models in text generation capabilities. (a) We evaluated five in-domain and three out-of-domain tasks using five report generation metrics, where FG, IG, FS, FG(Ind), and IG(Ind) denote finding generation, impression generation, finding summarization, finding generation with indication, and impression generation with indication, respectively, representing five distinct text generation tasks. (b) Sample data demonstration for the finding generation task. (c) Sample data demonstration for the impression generation task. (d) Sample data demonstration for the finding summarization task.

Writing radiology reports based on the interpretation of X-ray images is a labor-intensive and time-consuming repetitive task for clinicians. **CX-Mind**, through a four-stage progressive training approach, has achieved significant advancements in text generation capabilities. This progress contributes to alleviating clinicians' workload and enhancing the diagnostic efficiency of medical institutions. To evaluate **CX-Mind**'s text generation performance, we conducted assessments across five tasks: finding generation, impression generation, and finding summarization, among others. To ensure a comprehensive evaluation, we not only focused on traditional metrics for report generation, such as character-level matching (e.g., BLEU, ROUGE), but also emphasized semantic similarity between the generated outputs and gold-standard reports. Experimental results in Figure 6 (a) demonstrate that **CX-Mind** achieves SOTA performance across various text generation tasks. Specifically, in the finding generation task, which requires robust anomaly detection capabilities, GPT-4o exhibited commendable performance compared to other baseline models. However, **CX-Mind** outperformed GPT-4o, achieving a 1.6% higher BertScore, 7.6% higher BLEU, and an average 11.1% higher ROUGE score in the "Finding Generation" task. In the "Finding Generation (Indication)" task, **CX-Mind** surpassed GPT-4o by 3.6% in BertScore, 21.7% in BLEU, and an average of 22% in ROUGE scores. These results underscore the model's precise and reliable report generation capabilities. For the finding summarization task, **CX-Mind** outperformed the best-performing

baseline model, CheXagent, by an average of 2.2% across all evaluation metrics. Additionally, the impression generation task, which requires high-level summarization of radiological findings from images, revealed a substantial performance gap between most models and **CX-Mind**. Notably, **CX-Mind**'s generated impressions demonstrated high semantic consistency with ground-truth reports, achieving BertScores of 90.3% and 80.7% in the "Impression Generation" and "Impression Generation (Indication)" tasks, respectively. Furthermore, on an out-of-domain test set (OpenI), **CX-Mind** maintained SOTA performance in common report generation tasks, showcasing strong generalization capabilities.

#### 4.5.3. Performance on Spatiotemporal Alignment

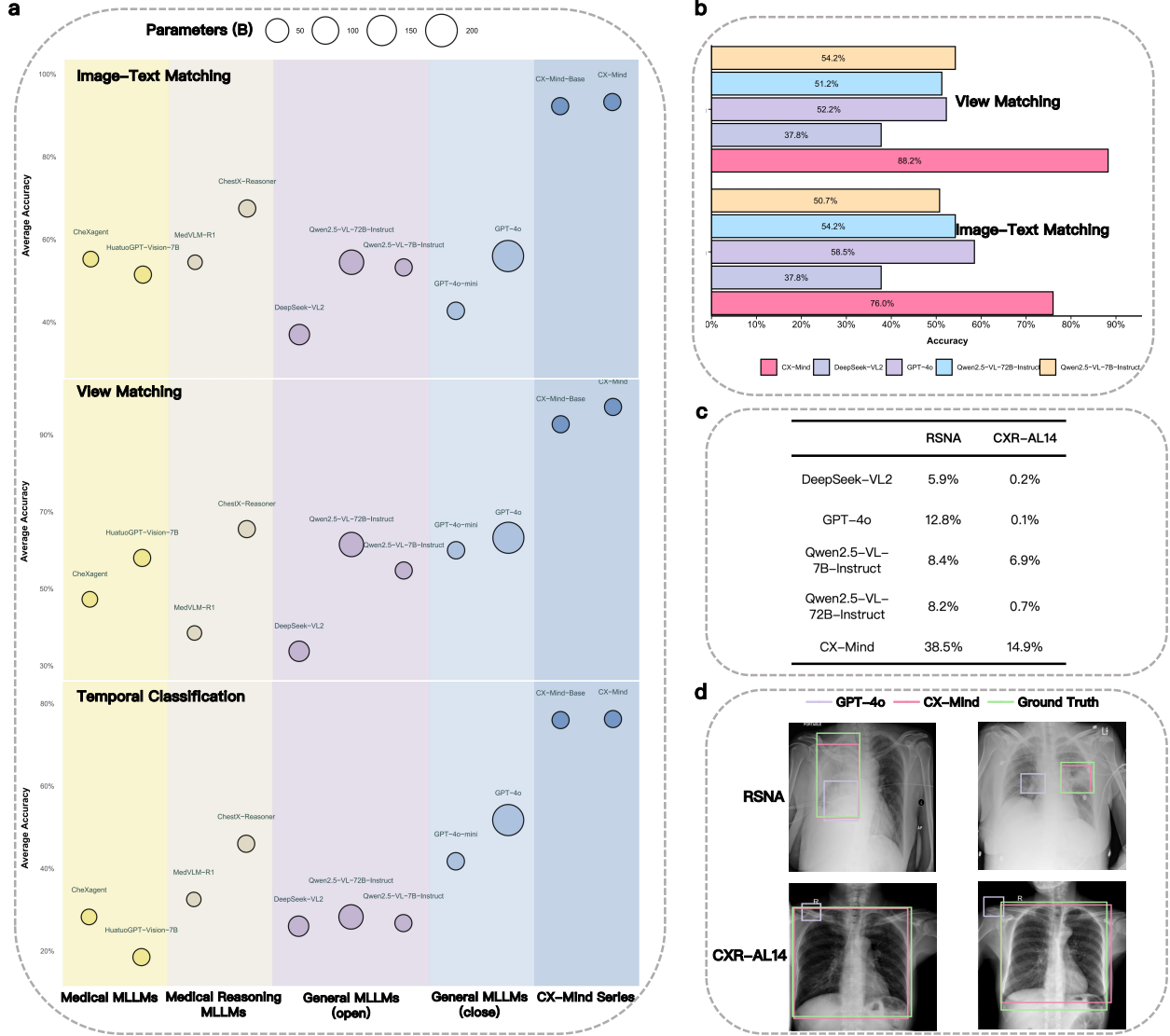


Figure 7: Comparative study of CX-Mind and various baseline models in spatiotemporal alignment capabilities. (a) Comparison of average precision between CX-Mind and four other types of MLLMs across image-text matching, view matching, and temporal classification tasks, where circle size represents model parameter scale. (b) Performance comparison of CX-Mind against mainstream models on an out-of-domain test set. (c) Comparison of IoU for different models in the disease target detection task. (d) Visualization results of disease localization range for each model.

In addition to the core capabilities outlined previously, an exceptional "X-ray expert" should possess robust temporal and spatial perception abilities. For temporal perception, we constructed two tasks—image-text matching and disease progression assessment—using the

Table 1: Comparative study of CX-Mind and various baseline models in foundational language capabilities. The upper half of the table includes LLMs with large parameter scales, while the lower half includes MLLMs based on the Qwen2.5-VL-7B-Instruct model.

Model	Medical Language Understanding		Medical Language Generation	Complex Medical Reasoning	Medical Safety & Ethics	Medical QA		Mean
	CHIP-CDN	CMeEE	IMCS-V2-MRG	DDx-basic	MedSafety	MedHG	Med-Exam	
o1-mini	85.10	50.50	76.50	83.00	27.50	38.20	52.90	59.10
Llama-3.3-70B-Instruct	98.60	48.80	91.00	84.20	66.90	71.10	66.80	75.34
GPT-4o	91.40	45.90	79.60	84.00	47.10	73.80	64.90	69.53
Qwen2.5-VL-7B-Instruct	88.67	39.86	86.56	75.89	18.00	40.21	16.33	52.22
CX-Mind (stage1)	92.00(+3.33)	46.92(+7.06)	88.86(+2.30)	83.24(+7.35)	64.00(+46.00)	62.94(+22.73)	64.67(+48.34)	71.80(+19.58)
CX-Mind	89.33(+0.66)	41.34(+1.48)	86.73(+0.17)	82.40(+6.51)	49.33(+31.33)	58.47(+18.26)	43.33(+27.00)	64.42(+12.20)

MIMIC-CXR and MS-CXR datasets. The image-text matching task involves selecting radiological examinations from the same patient at different time points to match the corresponding X-ray image. The disease progression task extracts two sequential radiology reports from the same patient, utilizing DeepSeek-V3 [14] to assess the progression of a specific critical disease, thereby evaluating the model’s temporal perception capabilities. For spatial perception, we evaluated the model through image view classification and disease detection-localization tasks. Experimental results in Figure 7 (a) reveal a substantial performance gap between baseline models and **CX-Mind** in temporal perception tasks. Our model outperformed the best-performing baseline, ChestX-Reasoner, by an average of 25.8% in image-text matching and 30.2% in disease progression tasks, demonstrating superior temporal perception capabilities. In the simpler image view classification task, **CX-Mind** achieved near-perfect performance across multiple datasets. Compared to high-performing closed-source models GPT-4o and GPT-4o-mini, **CX-Mind** surpassed them by 31% and 37.2%, respectively. As shown in Figure 7 (b), this significant advantage in spatiotemporal alignment was consistent in the out-of-domain OpenI test set, where **CX-Mind** achieved 76% and 88.3% performance in image-text matching and view classification tasks, respectively. Furthermore, **CX-Mind** exhibited strong target localization capabilities, enabling the detection of potential anomalies in X-ray images. When compared to SOTA models with target detection functionality, **CX-Mind** achieved 38.5% and 14.9% in IoU metrics (Figure 7 (d)) on the out-of-domain RSNA and CXR-AL14 datasets, respectively.

#### 4.5.4. Performance on Foundational Language Capabilities

Fine-tuning MLLMs for visual tasks often leads to a degradation of their language capabilities. Consequently, we not only focused on **CX-Mind**’s performance in chest X-ray tasks but also prioritized maintaining its original language proficiency. This consistency is critical for addressing complex clinical medical tasks. To this end, we collected seven publicly available text-only medical datasets, encompassing five key medical scenarios: language understanding, language generation, complex reasoning, safety and ethics, and knowledge-based question answering. Experimental results show that after the first stage of SFT on text-only medical data, **CX-Mind** (stage1) achieved performance comparable to closed-source models o1-mini and GPT-4o across most benchmarks. Notably, **CX-Mind** (stage1) outperformed these models by an average of 3.7% on CHIP-CDN, 10.8% on IMCS-V2-MRG, and 26.7% on MedSafety. Compared to the SOTA text-only large language model Llama-3.3-70B-Instruct, which has ten times the parameter count, **CX-Mind** (stage1) exhibited only slightly inferior performance. More importantly, **CX-Mind** maintained robust medical language capabilities even after X-ray knowledge injection and interleaved reasoning enhancement. Our findings indicate that **CX-Mind** outperformed the baseline model Qwen2.5-VL-7B-Instruct and the closed-source

reasoning model o1-mini by an average of 12.2% and 5.3% across all benchmarks, respectively, without significant loss of precision compared to **CX-Mind** (stage1). These results collectively demonstrate that **CX-Mind** successfully enhances visual reasoning capabilities while preserving strong foundational language abilities.

#### 4.6. Ablation

In this section, we first conduct a comprehensive ablation to disentangle the contribution of each training stage. Specifically, we isolate the four sequential modules and quantify their individual impact on downstream performance. Second, although SFT based on CoT data has recently become the cold-start recipe, the relative value of answer-only samples versus CoT samples remains under-explored for medical VLMS. We therefore compare both data regimes. Besides, we investigate how close-ended and open-ended samples shapes RL, revealing distinct gains that depend on task structure. Finally, we benchmark several RL variants that differ only in their reward definitions, confirming that a well-designed reward signal is pivotal to stable policy improvement.

##### 4.6.1. Training Stages

In order to fully demonstrate the effectiveness of CuRL-VPR, we compare the performance of SFT stages 1, 2, 3 and RL stages on three types of benchmarks, and the results are shown in Table 2. To avoid the possible impact of extra reward policies, this part of the ablation experiments all use RL based on answer-only samples.

**The effectiveness and necessity of knowledge injection in SFT.** As shown in Table 2, Stage 1 enhances language ability, the model’s text generation capability improves significantly without sacrificing visual understanding. For example, on the finding summary task, the average metric increases by 5.3%. After Stage 2 large-scale image fine-tuning (Stage 1+2), the model’s overall performance increases from 27.3% to 51.3%. Further mixing CoT data and answer-only data for SFT (Stage 1+2+3) pushes the model to continue improving across all three capabilities. Moreover, compared to the variant that applies RL without Stage 2 training (Stage 1+3+4), its overall performance drops significantly to 41.6%, even below the variant without RL. This demonstrates that large-scale image fine-tuning not only establishes a robust visual-semantic foundation but also provides a convergent initialization strategy for RL.

**Cold-start RL further improves performance.** RL is built upon a strong base model to comprehensively enhance the VLM’s interleaved reasoning capability. The Deepseek-R0 style RL (Stage 1+2+4) achieves an overall score of 54.0%, while with cold-start, CX-Mind-Base combined with RL further increases by 2.7%. This shows that cold-start examples provide a better initialization for policy optimization, accelerating convergence and improving final performance.

Table 2: Ablation Results on training stages across tasks (higher is better). "Base" refers to Qwen2.5-VL-7B-Instruct. "Stage" denotes the model after applying the corresponding training stage. CX-Mind-Base is the model trained through the full pipeline, where the RL stage uses only outcome (answer) rewards. "w/" denotes "with". "Ind." refers to Indication part in clinical report. "B1" and "B2" denote two benchmarks for the same task, distinguished by different question formulations. **Bold** highlights the best scores in each segment.

Task	Dataset	Metric	Base	Stage 1	Stage 1+2	Stage 1+2+3	Stage 1+2+4	Stage 1+3+4	CX-Mind -Base
<i>Visual Understanding</i>									
Disease Identification (open)	CheXpert	Accuracy	10.0	11.5	11.5	11.5	11.5	11.0	<b>12.0</b>
Disease Identification (open)		Jaccard	11.9	11.5	12.5	21.8	43.0	20.7	<b>43.9</b>
Disease Identification (open)	MIMIC-CXR	Accuracy	35.5	35.0	35.0	36.5	35.5	35.0	<b>37.5</b>

*Continued on next page*

Continued from previous page

Task	Dataset	Metric	Base	S1	S1+2	S1+2+3	S1+2+4	S1+3+4	CX-Mind
Disease Identification (open)		Jaccard	36.2	44.0	44.0	46.8	47.9	44.3	<b>51.5</b>
Disease Identification (binary)	CheXpert	Accuracy	46.0	59.0	<b>80.5</b>	78.0	74.5	64.5	<b>80.5</b>
Disease Identification (binary)	MIMIC-CXR	Accuracy	46.5	61.0	72.0	77.0	77.5	65.0	<b>78.0</b>
Disease Identification (single)	CheXpert	Accuracy	25.0	35.5	64.0	74.5	71.0	65.5	<b>75.0</b>
Disease Identification (single)	MIMIC-CXR	Accuracy	31.0	51.0	83.5	85.5	87.0	69.5	<b>88.0</b>
Disease Identification (multiple)	CheXpert	Accuracy	25.5	59.5	89.5	93.5	91.0	82.5	<b>95.0</b>
Disease Identification (multiple)	MIMIC-CXR	Accuracy	21.0	52.5	84.0	95.0	90.5	62.5	<b>98.0</b>
<i>Text Generation</i>									
Finding Generation	MIMIC-CXR	BERTScore	82.5	83.1	87.5	87.9	88.6	82.7	<b>88.7</b>
Finding Generation		BLEU	7.2	8.2	16.7	17.0	17.3	12.0	<b>19.7</b>
Finding Generation		ROUGE-1	22.4	22.0	37.9	38.4	38.8	31.6	<b>40.7</b>
Finding Generation		ROUGE-2	4.2	4.9	15.6	15.9	16.0	9.8	<b>16.1</b>
Finding Generation		ROUGE-L	8.0	8.3	<b>29.2</b>	27.4	29.1	15.2	28.5
Finding Generation w/ Ind.	MIMIC-CXR	BERTScore	84.6	84.8	82.3	88.4	89.1	85.3	<b>90.5</b>
Finding Generation w/ Ind.		BLEU	7.0	7.8	29.9	<b>31.6</b>	30.6	10.4	30.4
Finding Generation w/ Ind.		ROUGE-1	21.4	22.9	47.2	48.5	48.5	23.0	<b>49.7</b>
Finding Generation w/ Ind.		ROUGE-2	2.9	2.9	30.5	27.9	<b>33.5</b>	7.7	28.5
Finding Generation w/ Ind.		ROUGE-L	13.3	14.7	37.0	36.4	<b>37.4</b>	23.6	<b>37.4</b>
Finding Summary	MIMIC-CXR	BERTScore	85.5	85.6	92.9	93.3	93.4	87.7	<b>93.6</b>
Finding Summary		BLEU	7.5	8.4	45.3	46.1	46.8	30.1	<b>47.4</b>
Finding Summary		ROUGE-1	19.3	19.8	60.3	63.4	63.5	40.6	<b>65.1</b>
Finding Summary		ROUGE-2	9.1	9.9	48.6	50.9	51.9	30.7	<b>53.1</b>
Finding Summary		ROUGE-L	17.2	17.7	58.8	61.5	62.0	50.9	<b>63.4</b>
Impression Generation	MIMIC-CXR	BERTScore	83.6	83.4	88.0	88.4	89.1	84.1	<b>89.2</b>
Impression Generation		BLEU	3.7	5.7	26.6	27.2	26.8	6.8	<b>27.6</b>
Impression Generation		ROUGE-1	10.0	10.7	36.8	38.7	38.1	14.1	<b>44.5</b>
Impression Generation		ROUGE-2	2.0	2.2	27.2	28.2	27.8	15.7	<b>31.8</b>
Impression Generation		ROUGE-L	7.2	5.5	36.3	37.8	37.5	10.6	<b>42.8</b>
Impression Generation w/ Ind.	MIMIC-CXR	BERTScore	48.4	48.6	55.1	55.0	55.4	50.1	<b>78.3</b>
Impression Generation w/ Ind.		BLEU	2.3	5.3	18.1	17.4	18.1	17.4	<b>18.3</b>
Impression Generation w/ Ind.		ROUGE-1	7.0	7.8	20.5	19.1	21.0	13.9	<b>21.8</b>
Impression Generation w/ Ind.		ROUGE-2	1.1	2.0	13.9	14.2	<b>15.2</b>	7.7	<b>15.2</b>
Impression Generation w/ Ind.		ROUGE-L	5.9	7.1	20.3	19.1	<b>21.9</b>	10.5	21.5
<i>Spatiotemporal Alignment</i>									
Image-Text Match	B1	Accuracy	47.5	48.0	81.0	<b>92.0</b>	86.0	75.0	90.5
Image-Text Match	B2	Accuracy	59.0	60.0	92.5	92.5	<b>94.5</b>	81.5	94.0
Temporal Classification	MS-CXR	Accuracy	26.5	35.5	68.5	67.5	70.0	54.5	<b>71.5</b>
Temporal Classification	MIMIC-CXR	Accuracy	27.0	43.5	<b>80.5</b>	77.5	79.5	67.0	<b>80.5</b>
View Match	B1	Accuracy	48.5	59.5	78.0	<b>96.0</b>	87.0	71.0	92.0
View Match	B2	Accuracy	61.0	63.0	63.0	69.0	69.0	64.0	<b>93.5</b>
<b>Average</b>			27.3	31.9 <sup>+4.6</sup>	51.3 <sup>+24.0</sup>	53.5 <sup>+26.2</sup>	54.0 <sup>+26.7</sup>	41.6 <sup>+14.3</sup>	<b>56.7<sup>+29.4</sup></b>

#### 4.6.2. Training Data

To demonstrate the effectiveness of each part of the training data, we performed ablations on both the cold-start data and the curriculum-based RL. Similarly, to avoid interference, all models undergoing RL here use only the final-answer reward  $R_{\text{final}}$ .

**Cold-start with answer-only supervision samples is important.** To further validate the value of answer-only samples in cold-start and subsequent reinforcement learning, we also evaluated the performance when introducing answer-only data in Stage 3 (variant CX-Mind 1+2+3) and then applying RL on this basis (variant CX-Mind 1+2+3+4 (w/ Ans)). As shown in Figure 8, using only CoT for cold-start (CX-Mind 1+2+3 (w/o Ans)) leads to performance degradation on certain tasks; for example, finding generation drops by 1.61% compared to the Stage 2 baseline, and after adding RL (CX-Mind 1+2+3+4 (w/o Ans)), the average accuracy further decreases by 0.857% and the ROUGE-L score by 3.33% relative to CX-Mind-Base. In contrast, when we mix in answer-only samples during Stage 3, the model’s capabilities are significantly restored and improved.

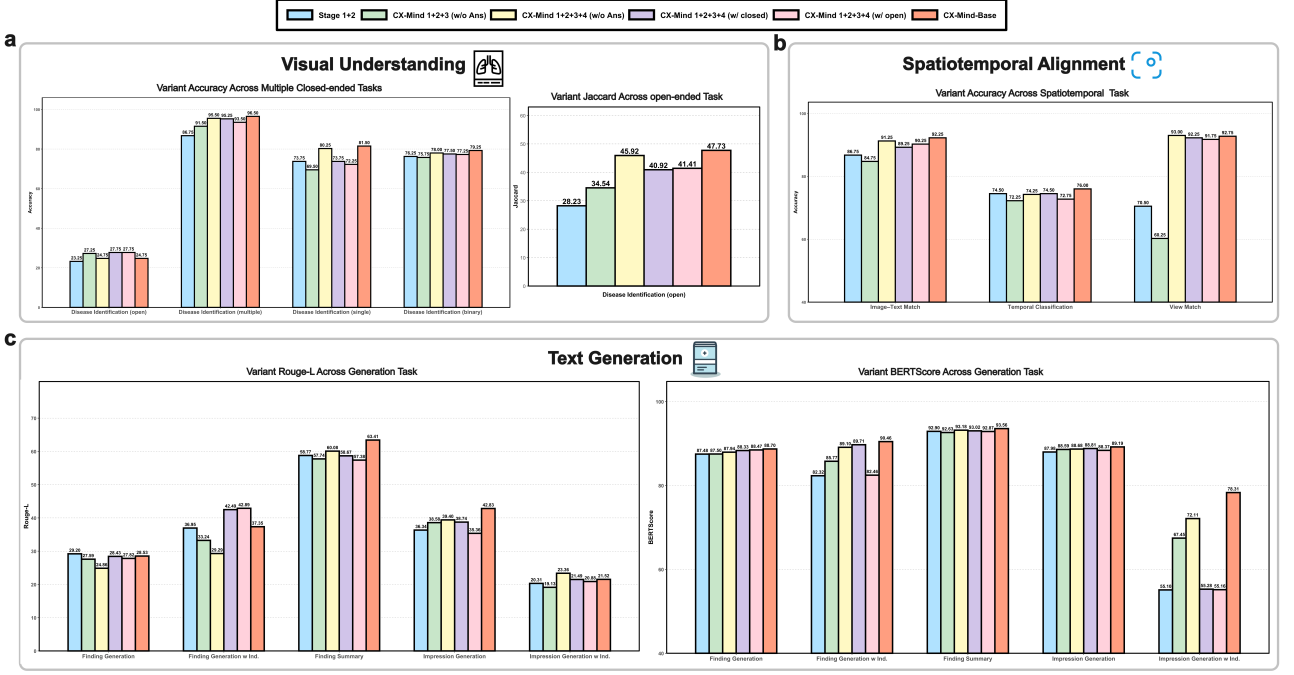


Figure 8: Ablation study on the impact of training data for our trained models, including **CX-Mind** stage1+2, **CX-Mind** 1+2+3 (w/o Ans), **CX-Mind** 1+2+3+4 (w/o Ans), **CX-Mind** 1+2+3+4 (w/ closed), **CX-Mind** 1+2+3+4 (w/ open), and **CX-Mind-Base**. **a.** Comparison of Visual Understanding capabilities among variants, covering five disease identification tasks evaluated using Accuracy and Jaccard metrics. **b.** Comparison of Visual Understanding capabilities among variants, including three localization and progression tasks. **c.** Comparison of Text Generation capabilities among variants, covering five generation tasks.

**Close-ended and open-ended RL further enhance reasoning capabilities.** Compared to CX-Mind-Base, omitting either close-ended or open-ended data during RL results in degraded performance. In visual understanding, disease identification Jaccard on open-ended examples falls by 6.32% without close-ended rewards and by 6.81% without open-ended rewards. In text generation, the Rouge-L for finding summary decreases by 6.03% when close-ended rewards are removed and by 4.74% when open-ended rewards are omitted. Spatiotemporal alignment average accuracy drops by 2.1%. Thus, both reward types are crucial for effective RL in interleaved cross-modal reasoning.

#### 4.6.3. Different Reward Strategies

We evaluate the effectiveness of different reward configurations in Table 3, using **CX-Mind-Base** trained solely on  $D_A$  as the baseline for comparison.

**Direct supervision at every step negatively impacts model performance.** Introducing direct step-wise thinking supervision (Direct Think) reduces the overall average metric from 56.7% to 54.1%, suggesting that unconstrained intermediate supervision may hinder effective model learning and exacerbate the credit assignment issue.

**Two additional process-based rewards are important.** Incorporating conditional thinking rewards (Base (Think)) slightly raises the average performance to 56.9% (+0.2%). When further adding process-answer rewards during close-ended RL (w/  $r_{\text{ans}}^{(\text{open})}$ ) and open-ended RL stages (w/  $r_{\text{ans}}^{(\text{close})}$ ), the average metric improves further. Finally, **CX-Mind**, integrating all reward strategies and mechanisms, achieves the best overall performance across tasks. These results highlight the necessity of process-oriented rewards guided by real medical reports for multimodal interleaved diagnostic reasoning in medical domains.

Table 3: Reward ablation across tasks (higher is better). "Direct Think" applies step-wise think rewards unconditionally at every training step. "Base (Think)" applies step-wise think rewards conditionally. In addition to (Think), "w/  $r_{\text{ans}}^{(\text{open})}$ " and "w/  $r_{\text{ans}}^{(\text{close})}$ " are variants that add step-wise answer rewards during the open-ended and close-ended RL stages, respectively. "CX-Mind" is the full model with both process and answer rewards. **Bold** highlights the best score in each row.

Task	Dataset	Metric	CX-Mind -Base	Direct Think	Base (Think)	w/ $r_{\text{ans}}^{(\text{open})}$	w/ $r_{\text{ans}}^{(\text{close})}$	CX-Mind
<i>Visual Understanding</i>								
Disease Identification (open)	CheXpert	Accuracy	12.0	11.5	12.0	12.0	11.5	<b>13.0</b>
Disease Identification (open)		Jaccard	43.9	11.5	42.0	40.3	39.3	<b>46.7</b>
Disease Identification (open)	MIMIC-CXR	Accuracy	37.5	<b>44.0</b>	40.0	41.0	40.5	41.5
Disease Identification (open)		Jaccard	51.5	44.0	51.0	52.0	51.5	<b>52.3</b>
Disease Identification (binary)	CheXpert	Accuracy	<b>80.5</b>	77.5	77.0	77.5	77.5	78.0
Disease Identification (binary)	MIMIC-CXR	Accuracy	78.0	77.5	78.5	78.5	79.5	<b>82.0</b>
Disease Identification (single)	CheXpert	Accuracy	75.0	77.5	79.5	<b>80.0</b>	73.0	77.5
Disease Identification (single)	MIMIC-CXR	Accuracy	88.0	81.0	85.5	<b>90.0</b>	86.0	87.0
Disease Identification (multiple)	CheXpert	Accuracy	95.0	94.5	94.5	<b>96.0</b>	94.5	95.5
Disease Identification (multiple)	MIMIC-CXR	Accuracy	<b>98.0</b>	96.0	96.0	96.5	97.0	97.0
<i>Text Generation</i>								
Finding Generation	MIMIC-CXR	BERTScore	<b>88.7</b>	88.1	88.1	87.9	87.6	87.8
Finding Generation		BLEU	<b>19.7</b>	15.4	18.6	18.4	18.0	18.7
Finding Generation		ROUGE-1	<b>40.7</b>	36.6	41.3	38.4	41.3	42.6
Finding Generation		ROUGE-2	<b>16.1</b>	13.2	15.9	16.2	15.5	16.9
Finding Generation		ROUGE-L	<b>28.5</b>	26.4	28.2	28.7	26.5	29.0
Finding Generation w/ Ind.	MIMIC-CXR	BERTScore	90.5	90.9	89.9	90.8	90.9	<b>91.0</b>
Finding Generation w/ Ind.		BLEU	30.4	<b>32.8</b>	33.9	33.4	33.3	34.2
Finding Generation w/ Ind.		ROUGE-1	49.7	51.4	48.9	51.2	52.1	<b>52.0</b>
Finding Generation w/ Ind.		ROUGE-2	28.5	31.6	32.5	33.6	<b>33.5</b>	33.0
Finding Generation w/ Ind.		ROUGE-L	37.4	41.3	39.3	41.7	<b>41.7</b>	41.8
Finding Summary	MIMIC-CXR	BERTScore	<b>93.6</b>	93.0	93.1	93.1	93.0	93.1
Finding Summary		BLEU	<b>47.4</b>	44.6	46.4	47.8	46.6	47.4
Finding Summary		ROUGE-1	<b>65.1</b>	61.7	65.9	65.2	60.9	66.1
Finding Summary		ROUGE-2	<b>53.1</b>	48.8	54.7	54.6	53.1	55.8
Finding Summary		ROUGE-L	<b>63.4</b>	60.0	63.7	64.4	62.2	64.3
Impression Generation	MIMIC-CXR	BERTScore	<b>89.2</b>	90.9	89.9	90.8	90.9	90.7
Impression Generation		BLEU	<b>27.6</b>	28.6	30.2	30.5	29.2	27.7
Impression Generation		ROUGE-1	44.5	38.3	44.8	45.0	45.5	<b>45.6</b>
Impression Generation		ROUGE-2	31.8	27.8	31.2	28.7	30.3	<b>32.7</b>
Impression Generation		ROUGE-L	42.8	37.5	43.0	43.2	40.5	<b>43.6</b>
Impression Generation w/ Ind.	MIMIC-CXR	BERTScore	<b>78.3</b>	59.8	71.2	76.2	78.8	80.7
Impression Generation w/ Ind.		BLEU	18.3	32.9	19.0	14.1	21.3	<b>22.7</b>
Impression Generation w/ Ind.		ROUGE-1	21.8	51.4	20.6	20.0	21.3	<b>21.9</b>
Impression Generation w/ Ind.		ROUGE-2	15.2	32.7	15.9	14.4	16.4	<b>17.4</b>
Impression Generation w/ Ind.		ROUGE-L	21.5	22.7	20.3	19.3	21.9	<b>21.9</b>
<i>Spatiotemporal Alignment</i>								
Image-Text Match	B1	Accuracy	90.5	91.5	90.5	88.5	90.0	<b>92.0</b>
Image-Text Match	B2	Accuracy	94.0	92.0	94.0	94.0	92.0	<b>94.5</b>
Temporal Classification	MS-CXR	Accuracy	71.5	71.5	71.0	71.0	<b>72.0</b>	72.0
Temporal Classification	MIMIC-CXR	Accuracy	80.5	80.0	<b>80.5</b>	80.0	81.0	80.5
View Match	B1	Accuracy	92.0	92.0	<b>95.0</b>	95.0	95.5	95.5
View Match	B2	Accuracy	93.5	96.5	98.5	98.5	<b>99.5</b>	99.0
<b>Average</b>			56.7	54.1 <sup>-2.6</sup>	56.9 <sup>+0.2</sup>	57.1 <sup>+0.4</sup>	56.8 <sup>+0.1</sup>	<b>58.2<sup>+1.5</sup></b>

#### 4.7. Clinical Evaluation

To comprehensively demonstrate **CX-Mind**'s outstanding performance, we conducted an extensive cross-center clinical evaluation. Specifically, the Rui-CXR dataset originates from 4,031 chest X-ray image reports collected at the Orthopedics Department of Ruijin Hospital, Shanghai Jiao Tong University, between 2018 and 2023. To protect patient privacy, all images and medical records were rigorously de-identified and approved by the hospital ethics committee. To prevent data leakage, all comparative models, including CX-Mind, were evaluated via local deployment.

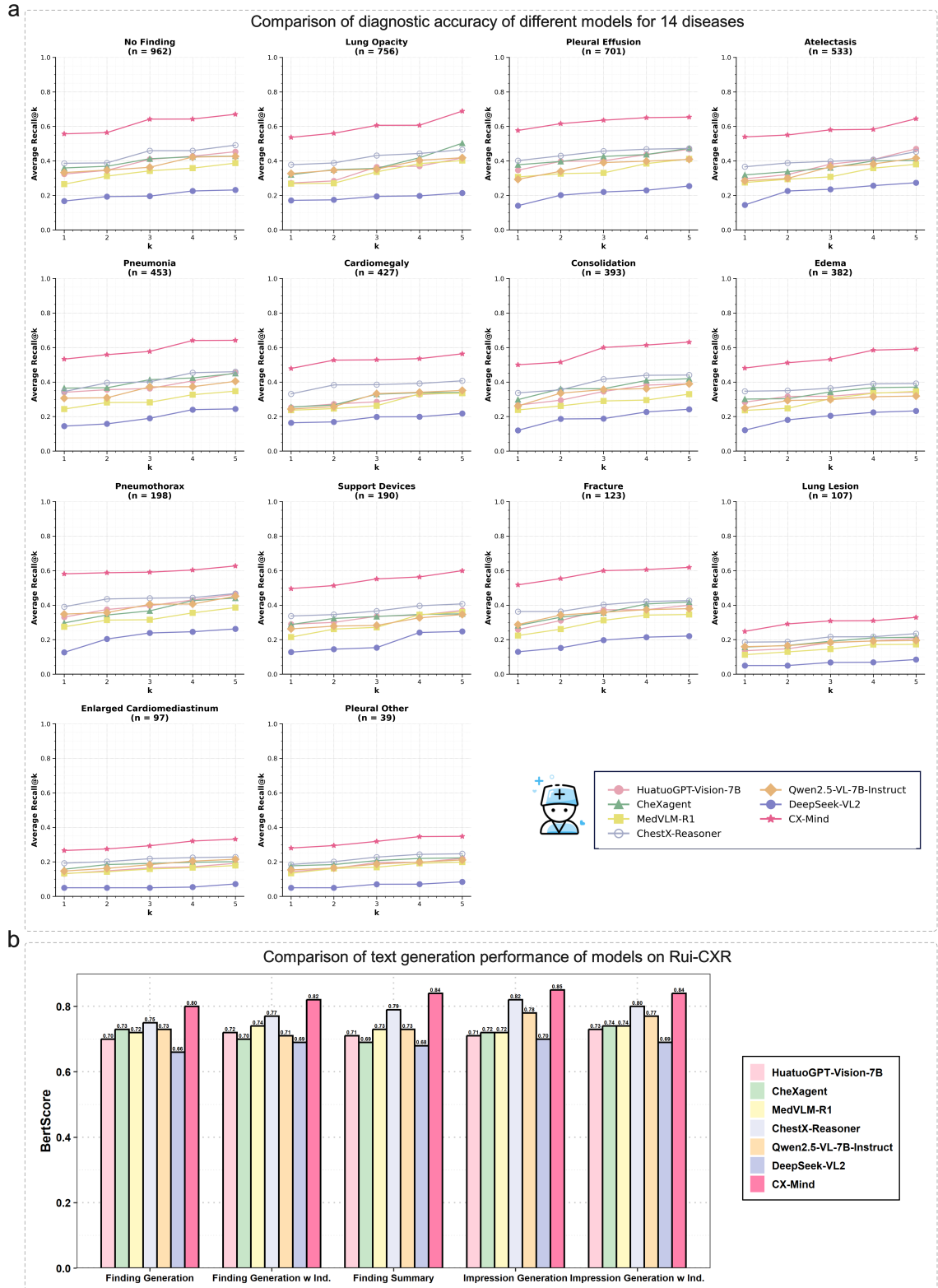


Figure 9: **Evaluation of different models on Rui-CXR.** (a) Diagnostic performance comparison of different models on 14 diseases. (b) Comparison of text generation performance of different models on Rui-CXR.

#### 4.7.1. Rui-CXR Dataset Curation

The Rui-CXR dataset initially included chest radiographs and accompanying text reports for 80,648 patients acquired by the standard posterior–anterior (PA) chest X-ray system at the Orthopedics Department of Ruijin Hospital, Shanghai Jiao Tong University, from 2018–2023. To align with the publicly available MIMIC-CXR dataset, this study focused on 14 common chest diseases (e.g., lung opacity, pleural effusion, Atelectasis, etc.) featured in MIMIC-CXR. We adopted a three-stage data filtering strategy: first, we performed an image quality screening on the original 80,648 images, excluding 12,292 samples with severe artifacts or non-standard PA views. Next, we applied automated keyword matching on the remaining images and reports to select 8,059 candidate images containing descriptions of the target lesions. Based on report completeness checks and diagnostic label consistency verification, we further filtered out 4,028 samples with overly brief reports or mismatched labels, resulting in a high-quality, fully de-identified test set of 4,031 images. Since patients may present multiple concurrent diseases, the total count of disease annotations exceeds 4,031.

Based on these 4,031 high-quality images, we constructed two real-world clinical evaluation tasks: (1) a disease diagnosis task based on recall@ $k$ , where we prompt the model to output  $k$  candidate diagnostic labels and quantify the model’s coverage by computing the proportion of true labels in the  $k$  predictions. (2) an open-ended report generation task, in which the model, given a single chest X-ray input, generates a radiology report conforming to clinical writing standards, and its accuracy and completeness are assessed using quantitative metrics.

#### 4.7.2. Evaluation Result

As shown in Figure 9 (a), CX-Mind demonstrates remarkable diagnostic capabilities on Rui-CXR: its top-1 diagnostic accuracy for lung opacity reaches 0.53, significantly outperforming comparative models by 0.20 on average, and achieves the best performance among common diseases such as pleural effusion and Atelectasis, fully validating its high coverage in multi-candidate diagnostic scenarios. Accuracy declines for conditions such as lung lesion and pleural other, indicating areas for clinical improvement.

As illustrated in Figure 9(b), CX-Mind secures the highest scores across all five text generation tasks: standard Finding Generation BERTScore 0.80 (w/ Ind. 0.82), averaging a 5% gain over the second-best models. These results convincingly demonstrate **CX-Mind**’s exceptional ability in fusing chest X-ray imaging with clinical text for reasoning and high-quality report generation.

#### 4.7.3. Expert Assessment

We invited internists and surgeons from multiple medical centers to conduct a subjective evaluation of 390 chest radiographs drawn from the Rui-CXR dataset, covering 13 common thoracic diseases (excluding the "No Finding" category, 30 cases per disease). The diagnostic outputs of Qwen2.5-VL-7B-Instruct, CheXagent, ChestX-Reasoner, and CX-Mind were compared. All cases were fully de-identified, and both model names and any potentially identifying information were masked. The images and model responses were randomly shuffled, yielding a total of 1,560 independent evaluation samples.

For each case, physicians scored the model answer on five dimensions:

- CRS (Clinical Relevance Score): whether the conclusion matches clinical priorities, terminology conventions, and management recommendations.
- LCS (Logical Coherence Score): whether the reasoning chain is complete and self-consistent, free of hallucinations and logical jumps.
- ESS (Evidence Support Score): whether key image or textual evidence is sufficiently and accurately cited and directly supports the conclusion.

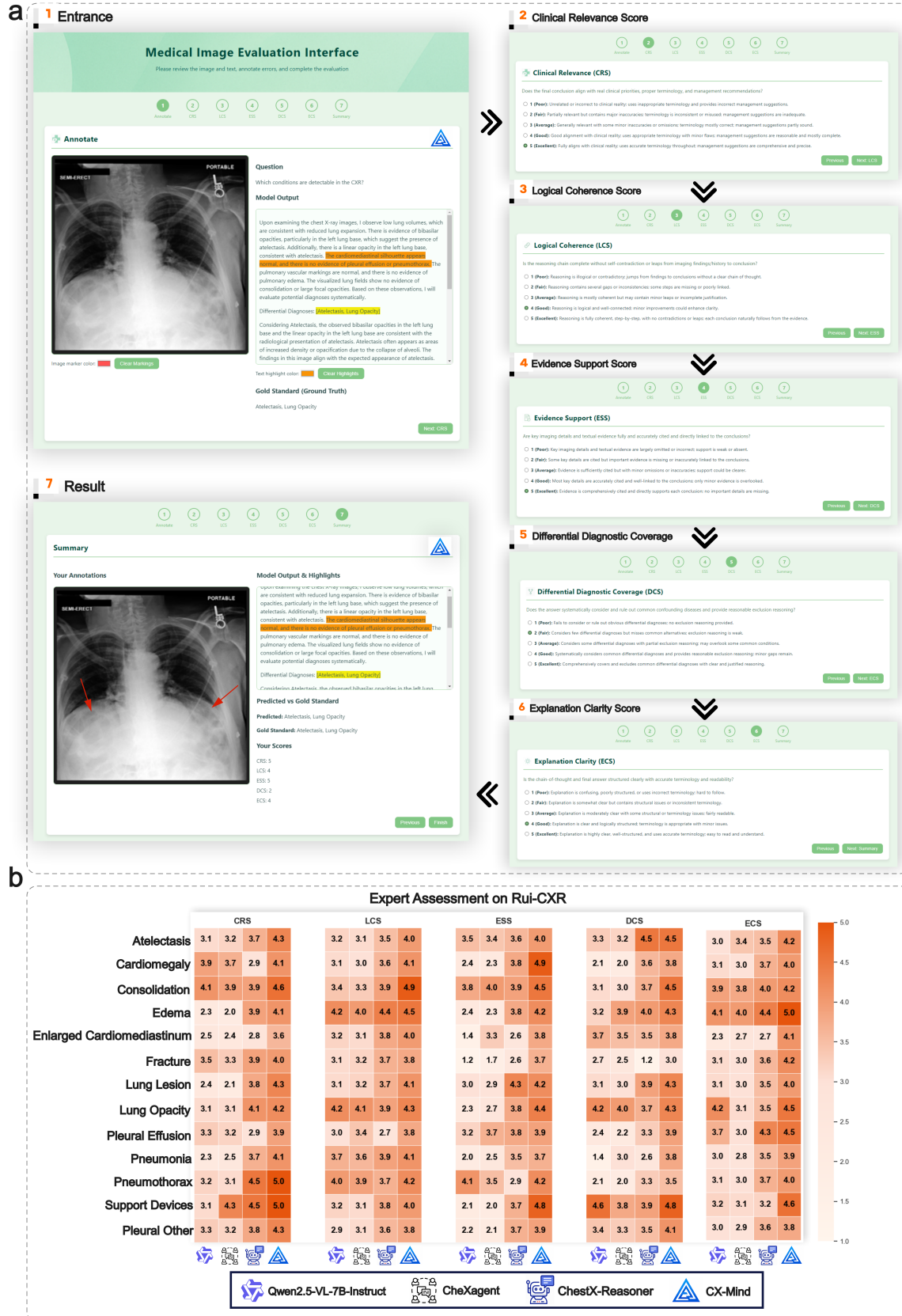


Figure 10: **Expert evaluation of CX-Mind and baselines by multi-center clinicians.** (a) Physician scoring workflow interface: the system randomly presents a de-identified chest radiograph and an anonymized model output. Following unified guidelines, physicians annotate the image and text, then rate CRS, LCS, ESS, DCS, and ECS in sequence. (b) Five-dimension expert-score heatmap on 13 thoracic diseases from Rui-CXR. CX-Mind attains the highest average score on all dimensions.

- DCS (Differential Diagnostic Coverage): whether common differential diagnoses are systematically considered and reasonably ruled out.
- ECS (Explanation Clarity Score): clarity of structure, terminology accuracy, and readability of both the reasoning process and the final answer.

To facilitate efficient and convenient assessment, we developed a user-friendly web interface in Figure 10(a). The system randomly presents a de-identified chest radiograph and an anonymous model output. After reading unified guidelines, physicians freely annotate the image (marking suspicious findings) and the text (highlighting inappropriate statements), compare them with the displayed ground-truth answer, and then complete the above five ratings on a 1–5 scale (1 = poor, 5 = excellent). Approximately 10% of the cases are re-sampled to check intra-rater consistency.

CX-Mind consistently achieved the highest scores across CRS, LCS, ESS, DCS, and ECS, demonstrating comprehensive superiority from evidence citation to reasoning and final wording. ChestX-Reasoner ranked at an intermediate level (around 3.5 on average). Its logic and evidence were acceptable, but differential coverage and clarity lagged behind. CheXagent and Qwen2.5-VL-7B-Instruct received lower average scores. Notably, clinicians strongly preferred **CX-Mind**’s interleaved reasoning paradigm: they could directly inspect the thought process, rapidly judge its soundness, and intervene promptly when necessary, thereby shortening the feedback loop and enhancing trust and adoption of the model’s output.

#### 4.8. Qualitative analysis

Figure 11 presents a visual comparison of **CX-Mind**’s performance in the co-morbidity identification task against the closed-source model GPT-4o and the comparable model CheXagent. The content between green and pink tags represents **CX-Mind**’s reasoning process and intermediate answers (final answer), respectively, while orange annotations highlight information that is inconsistent with medical facts or deviates from clinical observations. The figure clearly demonstrates that both large-scale closed-source models and downstream foundation models fine-tuned on specific X-ray corpora struggle to avoid hallucinations. Such erroneous or fabricated information hinders the practical application of MLLMs in clinical settings. Specifically, in Figure 11 (a), CheXagent attributes lung opacity to potential infection or inflammation, yet the X-ray images lack supporting evidence for these causes (e.g., no typical signs of infection or inflammation, such as pulmonary consolidation). This results in a diagnosis lacking specificity, as it fails to systematically rule out conditions based on available imaging evidence. Similarly, GPT-4o’s reasoning process suggests that increased lung markings may be associated with pulmonary edema. However, this observation alone is insufficient to confirm pulmonary edema, as other conditions (e.g., chronic lung disease) may present similar findings. This process lacks further differential diagnosis to enhance the reliability of the results. In contrast, **CX-Mind** adheres to our predefined reasoning paradigm, producing a meticulous and accurate interleaved reasoning process. This improves its interactivity with clinicians and enhances its potential for deployment in real-world diagnostic scenarios.

## 5. Conclusion & Discussion

We have presented CX-Mind, the first generative chest X-ray diagnostic model that integrates an interleaved reasoning paradigm with structured process rewards within a curriculum learning reinforcement learning framework, achieving a balance between clinical performance and interpretability. The training process follows a staged curriculum in which the model first acquires foundational knowledge via supervised fine-tuning, then transitions into reinforcement learning with GRPO—beginning with close-ended tasks to establish core reasoning skills before

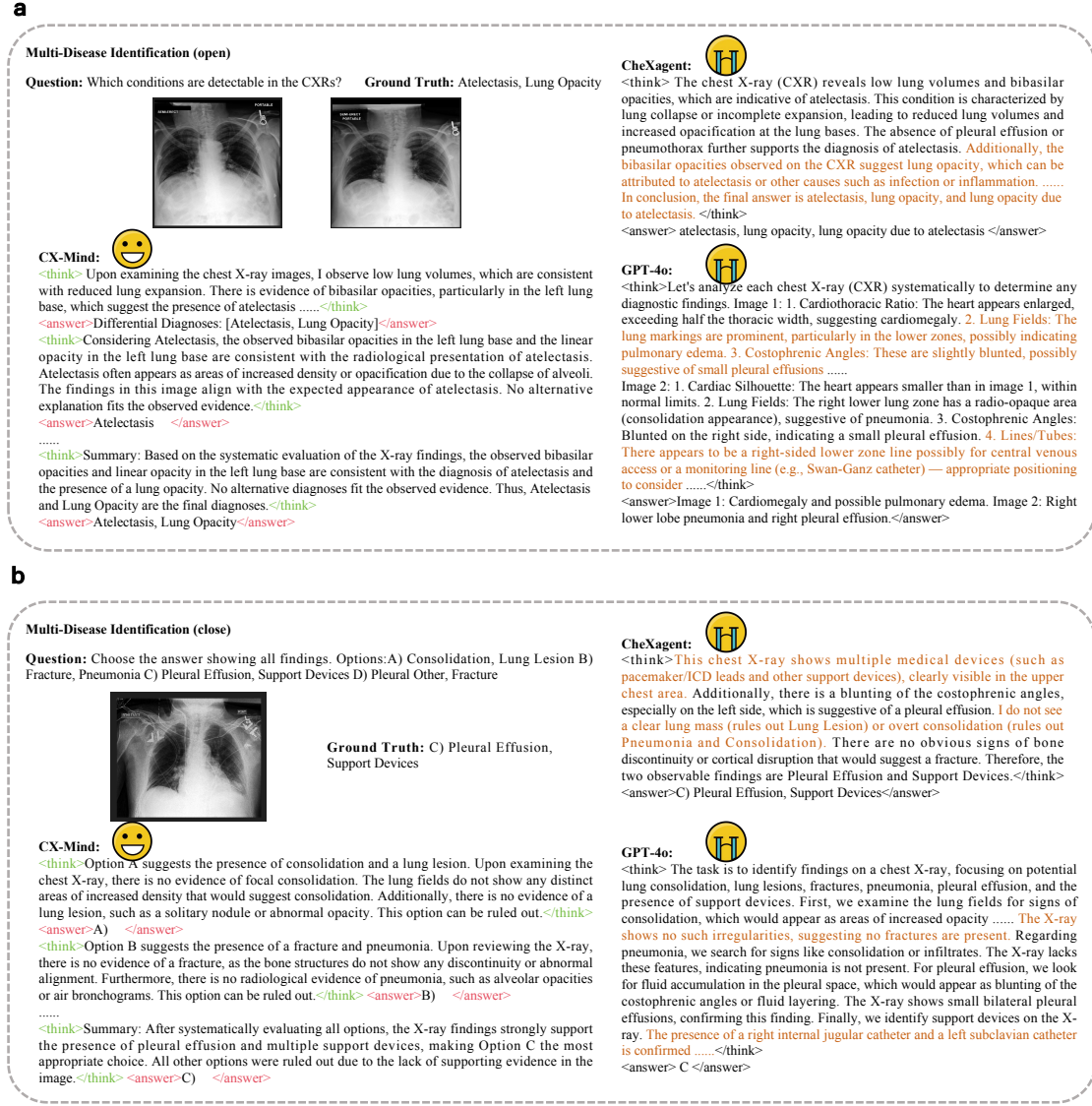


Figure 11: **Comparative examples of diagnostic responses from GPT-4o, CheXagent and CX-Mind.** (a) Open-ended diagnostic question results. (b) Close-ended diagnostic question results.

advancing to open-ended diagnostic queries. By interleaving concise thinking and answering steps, CX-Mind mirrors the diagnostic workflow of radiologists, allowing each intermediate conclusion to be visible and verifiable. This design not only improves transparency but also enables clinicians to intervene if a reasoning path deviates, thereby reducing the risk of compounding errors. The process reward mechanism, which grants additional feedback only when both format and accuracy criteria are met, encourages the model to refine its reasoning without overwhelming it with premature supervision.

Extensive evaluations demonstrate that CX-Mind outperforms a broad spectrum of SOTA baselines across disease identification, report generation, and spatiotemporal alignment tasks. In particular, the interleaved outputs yield more focused and succinct chains of thought compared to traditional chain-of-thought methods, which often produce verbose and redundant reasoning. Our ablation studies further confirm that the synergy between interleaved reasoning, curriculum reinforcement learning, and process-based rewards is critical: without this combined framework, gains in final accuracy come at the expense of coherent intermediate steps, whereas CX-Mind achieves consistent improvements in both performance metrics and clinician trust.

Despite these advances, real-world deployment will require careful consideration of several

factors. The reliance on publicly available datasets may not fully capture the heterogeneity of clinical imaging protocols or rare pathologies encountered in practice. Moreover, process rewards depend on ground-truth annotations that might not reflect every nuance of expert radiological judgment. Computational demands for generating and evaluating interleaved reasoning sequences also present challenges for integration into time-sensitive clinical workflows.

Future work should explore adaptive reward schedules that respond to task complexity and incorporate human feedback loops for on-the-fly calibration. Extending the interleaved paradigm and curriculum RL framework to additional imaging modalities and integrating patient history or laboratory findings could further enhance clinical utility. By grounding model outputs in structured, inspectable steps and progressing through a carefully designed curriculum, CX-Mind paves the way for AI systems that collaborate seamlessly with healthcare professionals, ultimately improving diagnostic accuracy and patient outcomes.

### **CRedit authorship contribution statement**

**Wenjie Li:** Conceptualization, Methodology, Investigation, Formal analysis, Visualization, Writing - original draft, Writing - review & editing. **Yujie Zhang:** Software, Data curation, Formal analysis, Validation, Writing - review & editing. **Haoran Sun:** Methodology, Validation, Software, Formal analysis, Writing - review & editing. **Yueqi Li:** Investigation, Data curation, Writing - review & editing. **Fanrui Zhang:** Software, Visualization. **Victoria Borja Clausich:** Investigation. **Sade Mellin:** Investigation. **Renhao Yang:** Resources, Investigation. **Chenrun Wang:** Formal analysis, Data curation, Visualization. **Zih-Shuo Wang:** Data curation. **Shiyi Yao:** Investigation. **Gen Li:** Investigation. **Yidong Xu:** Investigation, Data curation. **Hanyu Wang:** Software, Data curation. **Yilin Huang:** Data curation. **Angela Lin Wang:** Writing - review & editing. **Chen Shi:** Resources, Investigation. **Yin Zhang:** Validation, Supervision. **Jianan Guo:** Software. **Luqi Yang:** Data curation. **Renxuan Li:** Formal analysis. **Yang Xu:** Validation, Investigation. **Jiawei Liu:** Software. **Lei Liu:** Methodology, Validation, Writing - review & editing. **Carlos Gutiérrez SanRomán:** Investigation, Resources. **Lei Wang:** Conceptualization; Supervision, Project administration, Funding acquisition, Writing - review & editing.

### **Declaration of competing interest**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### **Acknowledgments**

The paper was funded by Key Research and Development Plan of the Ministry of Science and Technology (2023YFC2410705), the National Natural Science Foundation of China (32171317 and 82472419) and Shanghai Municipal Science and Technology Commission Project (25ZR1402348).

### **Data Availability**

Public Data, the preprocess code will be available on request.

## References

- [1] A. Jaech, A. Kalai, A. Lerer, A. Richardson, A. El-Kishky, A. Low, A. Helyar, A. Madry, A. Beutel, A. Carney, et al., Openai o1 system card, arXiv preprint arXiv:2412.16720 (2024).
- [2] T. Doshi, Gemini 2.5: Our most intelligent models are getting even better, Recuperado el 15 (2025).
- [3] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, et al., Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, arXiv preprint arXiv:2501.12948 (2025).
- [4] Y. Wang, S. Wu, Y. Zhang, S. Yan, Z. Liu, J. Luo, H. Fei, Multimodal chain-of-thought reasoning: A comprehensive survey, arXiv preprint arXiv:2503.12605 (2025).
- [5] A. Holzinger, M. Dehmer, F. Emmert-Streib, R. Cucchiara, I. Augenstein, J. Del Ser, W. Samek, I. Jurisica, N. Díaz-Rodríguez, Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence, *Information Fusion* 79 (2022) 263–278.
- [6] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, F. Herrera, Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence, *Information fusion* 99 (2023) 101805.
- [7] Z. Chen, M. Varma, J. Xu, M. Paschali, D. Van Veen, A. Johnston, A. Youssef, L. Blanke-meier, C. Bluethgen, S. Altmayer, et al., A vision-language foundation model to enhance efficiency of chest x-ray interpretation, arXiv e-prints (2024) arXiv–2401.
- [8] A. Fallahpour, J. Ma, A. Munim, H. Lyu, B. Wang, Medrax: Medical reasoning agent for chest x-ray, arXiv preprint arXiv:2502.02673 (2025).
- [9] Y. Lai, J. Zhong, M. Li, S. Zhao, X. Yang, Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models, arXiv preprint arXiv:2503.13939 (2025).
- [10] J. Pan, C. Liu, J. Wu, F. Liu, J. Zhu, H. B. Li, C. Chen, C. Ouyang, D. Rueckert, Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning, arXiv preprint arXiv:2502.19634 (2025).
- [11] N. Sharma, Cxr-agent: Vision-language models for chest x-ray interpretation with uncertainty aware radiology reporting, arXiv preprint arXiv:2407.08811 (2024).
- [12] A. S. Albahri, A. M. Duhaim, M. A. Fadhel, A. Alnoor, N. S. Baqer, L. Alzubaidi, O. S. Albahri, A. H. Alamoodi, J. Bai, A. Salhi, et al., A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion, *Information Fusion* 96 (2023) 156–191.
- [13] D. Xu, Y. Chen, Z. Chai, Y. Xiao, Y. Yan, W. Ding, H. Wang, Z. Jin, W. Jiao, W. Yue, et al., Knowledge fusion in deep learning-based medical vision-language models: A review, *Information Fusion* (2025) 103455.
- [14] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, et al., Deepseek-v3 technical report, arXiv preprint arXiv:2412.19437 (2024).

- [15] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, et al., Gpt-4o system card, arXiv preprint arXiv:2410.21276 (2024).
- [16] S. Rui, K. Chen, W. Ma, X. Wang, Improving medical reasoning with curriculum-aware reinforcement learning, arXiv preprint arXiv:2505.19213 (2025).
- [17] C. Bluethgen, P. Chambon, J.-B. Delbrouck, R. van der Sluijs, M. Połacin, J. M. Zambrano Chaves, T. M. Abraham, S. Purohit, C. P. Langlotz, A. S. Chaudhari, A vision-language foundation model for the generation of realistic chest x-ray images, *Nature Biomedical Engineering* (2024) 1–13.
- [18] D. Ma, J. Pang, M. B. Gotway, J. Liang, A fully open ai foundation model applied to chest radiography, *Nature* (2025) 1–11.
- [19] J. Yao, X. Wang, Y. Song, H. Zhao, J. Ma, Y. Chen, W. Liu, B. Wang, Eva-x: A foundation model for general chest x-ray analysis with self-supervised learning, arXiv preprint arXiv:2405.05237 (2024).
- [20] Z. Chen, M. Varma, J.-B. Delbrouck, M. Paschali, L. Blankemeier, D. Van Veen, J. M. J. Valanarasu, A. Youssef, J. P. Cohen, E. P. Reis, et al., Chexagent: Towards a foundation model for chest x-ray interpretation, arXiv preprint arXiv:2401.12208 (2024).
- [21] R. W. Lee, K. H. Lee, J. S. Yun, M. S. Kim, H. S. Choi, Comparative analysis of m4cxr, an llm-based chest x-ray report generation model, and chatgpt in radiological interpretation, *Journal of Clinical Medicine* 13 (23) (2024) 7057.
- [22] J. Chen, C. Gui, R. Ouyang, A. Gao, S. Chen, G. H. Chen, X. Wang, R. Zhang, Z. Cai, K. Ji, et al., Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale, arXiv preprint arXiv:2406.19280 (2024).
- [23] B. Wang, H. Zhao, H. Zhou, L. Song, M. Xu, W. Cheng, X. Zeng, Y. Zhang, Y. Huo, Z. Wang, et al., Baichuan-m1: Pushing the medical capability of large language models, arXiv preprint arXiv:2502.12671 (2025).
- [24] W. Zhang, P. Zhang, J. Guo, T. Cheng, J. Chen, S. Zhang, Z. Zhang, Y. Yi, H. Bu, Patho-r1: A multimodal reinforcement learning-based pathology expert reasoner, arXiv preprint arXiv:2505.11404 (2025).
- [25] T. Lin, W. Zhang, S. Li, Y. Yuan, B. Yu, H. Li, W. He, H. Jiang, M. Li, X. Song, et al., Healthgpt: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation, arXiv preprint arXiv:2502.09838 (2025).
- [26] Y. Luo, J. Zhang, S. Fan, K. Yang, M. Hong, Y. Wu, M. Qiao, Z. Nie, Biomedgpt: An open multimodal large language model for biomedicine, *IEEE Journal of Biomedical and Health Informatics* (2024).
- [27] R. Xie, D. Qiu, D. Gopinath, D. Lin, Y. Sun, C. Wang, S. Potdar, B. Dhingra, Interleaved reasoning for large language models via reinforcement learning, arXiv preprint arXiv:2505.19640 (2025).
- [28] Y. Bengio, J. Louradour, R. Collobert, J. Weston, Curriculum learning, in: *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 41–48.

- [29] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, S. Horng, Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports, *Scientific data* 6 (1) (2019) 317.
- [30] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, et al., Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, in: *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33, 2019, pp. 590–597.
- [31] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, C. J. McDonald, Preparing a collection of radiology examinations for distribution and retrieval, *Journal of the American Medical Informatics Association* 23 (2) (2016) 304–310.
- [32] M. Liu, W. Hu, J. Ding, J. Xu, X. Li, L. Zhu, Z. Bai, X. Shi, B. Wang, H. Song, et al., Medbench: A comprehensive, standardized, and reliable benchmarking system for evaluating chinese medical large language models, *Big Data Mining and Analytics* 7 (4) (2024) 1116–1128.
- [33] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, et al., Qwen2. 5-vl technical report, *arXiv preprint arXiv:2502.13923* (2025).
- [34] Z. Wu, X. Chen, Z. Pan, X. Liu, W. Liu, D. Dai, H. Gao, Y. Ma, C. Wu, B. Wang, et al., Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, *arXiv preprint arXiv:2412.10302* (2024).
- [35] Z. Fan, C. Liang, C. Wu, Y. Zhang, Y. Wang, W. Xie, Chestx-reasoner: Advancing radiology foundation models with reasoning through step-by-step verification, *arXiv preprint arXiv:2504.20930* (2025).