Teach-to-Reason: Competition-Guided Reasoning with a Self-Improving Teacher

Xiao Han **Hao Liu**[†,‡] **Zhimin Bao** **Jile Jiao**
Yue Wang **Hui Guo** **Xiaofeng Mou** **Yi Xu**[†]

AIRC, Midea Group
{hanxiao36,liuhao249,baozm8,jiaojl5}@midea.com
{wangyue184,guohui55,mouxf,xuyi42}@midea.com

[†]Corresponding authors [‡]Project leader

Abstract

Chest X-ray visual question answering (CXR VQA) requires models not only to predict correct answers, but also to produce reliable medical reasoning. However, existing reinforcement-learning-based training typically relies on answer-level rewards, which are often too coarse to improve chain-of-thought (CoT) quality and can become ineffective when group-level advantages collapse to zero. We propose **Teach-to-Reason (T2R)**, a framework that introduces comparison-based supervision into CoT optimization through a self-improving *Teacher* and a competition-guided *Reasoner*. As the Teacher is iteratively strengthened via self-competition, the Reasoner is optimized against progressively stronger Teacher-generated references. We further introduce a case-wise reward design that preserves the original reward-induced positive/negative partition when it is informative, and restores supervision from competition scores when the original reward signal degenerates. Experiments on multiple CXR open-ended VQA benchmarks show that T2R outperforms strong baselines overall, indicating that comparison-based supervision, when integrated in a controlled and principled manner, provides a more effective training signal for reasoning optimization. Our code is available at `https://github.com/midea-ai/Teach2Reason`.

1 Introduction

Chest X-ray visual question answering (CXR VQA) aims to answer clinically relevant questions grounded in chest radiographs, and has become an important testbed for medical multimodal understanding and reasoning [1–4]. Unlike general-domain VQA, CXR VQA requires not only correct answers, but also *reliable medical rationales* that justify those answers. In clinical settings, an apparently correct prediction may still rely on fragile, accidental, or clinically implausible reasoning. As a result, improving intermediate reasoning quality—rather than optimizing answer accuracy alone—is a central challenge in CXR VQA [5, 6].

Recent reinforcement-learning-based approaches have shown promise in improving reasoning ability in LLMs, but existing RL/RLVR-style training typically relies on *answer-level* rewards, such as whether the final answer is correct or satisfies a verifiable constraint [7–11]. Such signals are effective for optimizing final outcomes, yet they provide only weak supervision on the quality of the chain of thought (CoT) [6] itself. In particular, answer-level rewards are often too coarse to continuously shape reasoning quality, and when task rewards become identical within a group, the corresponding group-level advantage can degenerate to zero, making the training signal sparse or ineffective. A natural alternative is to directly reward CoTs using fixed rules [12] or an LLM-as-a-Judge [13], but

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

these approaches depend on predefined criteria and can quickly become sparse once the model learns to satisfy them [14, 15].

Motivated by these limitations, we propose **Teach-to-Reason (T2R)**, a framework that introduces comparison-based supervision into CoT optimization for CXR VQA through an iteratively improved *Teacher* and a competition-guided *Reasoner*. In T2R, the Teacher generates reference CoTs, while the Reasoner is the model used for final VQA inference. The Teacher is improved stage by stage through *self-competition*, and the Reasoner is optimized against progressively stronger Teacher-generated references. Compared with rule-based judge rewards, this design continually refreshes the supervision target as the Teacher evolves, making comparison-based supervision more adaptive and less susceptible to early saturation [16–18].

A second key ingredient of T2R is its reward design. We do not want additional CoT supervision to overwrite or destabilize the original positive/negative partition induced by the task reward within each GRPO group. Instead, T2R follows a case-wise principle: when the task reward already induces a meaningful partition, we preserve it and use competition scores only to refine the relative ordering within each side of the partition; when the task reward becomes completely uninformative and the group-level advantage degenerates to zero, we use competition scores to recover the partition and directly construct the final training signal. In this way, T2R preserves the original reward-induced preference whenever it is informative, while restoring supervision when the original signal collapses. Thus, T2R does not replace the task reward with competition scores; rather, it augments the original partition with finer-grained and more persistent supervision.

We evaluate T2R on Qwen3-VL-Instruct models at the 2B and 4B scales [19] across multiple CXR open-ended VQA benchmarks. The results show that T2R outperforms strong baselines overall. In summary, our main contributions are as follows:

- We propose **Teach-to-Reason (T2R)**, a training framework for CXR VQA that combines *iterative Teacher self-competition* with *competition-guided Reasoner optimization* to improve CoT quality.
- We introduce a case-wise reward design that preserves the original positive/negative partition induced by task rewards within each group, uses competition scores only for within-partition refinement, and recovers supervision when the task reward becomes degenerate.
- Extensive experiments on multiple CXR open-ended VQA benchmarks demonstrate that T2R outperforms strong baselines, including RLVR, Frozen Teacher, and Judge Reward, and is accompanied by denser and more persistent comparison-based supervision during training.

2 Related Work

CXR and Medical VQA. Medical visual question answering has been studied through benchmarks such as VQA-RAD [1], VQA-Med [20], PathVQA [21], and SLAKE [2], and more recently through larger chest-radiology resources such as MIMIC-Ext-MIMIC-CXR-VQA [3] and RexVQA [4]. Prior work has also explored stronger medical VQA architectures, multimodal pretraining, and medical multimodal foundation models, including MMBERT [22], M^{3}AE [23], BiomedCLIP [24], LLaVA-Med [25], Med-Flamingo [26], and MedGemma [27]. Together, these efforts have substantially advanced medical VQA benchmarks, model architectures, and multimodal medical modeling [28]. However, they have mainly focused on answer prediction, dataset construction, or general medical multimodal modeling, while the explicit optimization of chain-of-thought (CoT) quality in CXR VQA remains relatively underexplored.

Reasoning Optimization with RL and Judge-Based Supervision. Recent post-training methods have improved reasoning through preference optimization and RL/RLVR-style training [7–11]. Yet these approaches often rely on answer- or outcome-level supervision, which is typically too coarse to shape intermediate CoT quality [6, 29, 30]. A more direct line of work supervises reasoning with rubrics [12, 31–35], LLM-as-a-Judge signals [14, 13], or self-rewarding schemes [15]. However, such methods still assign scores to individual CoTs through predefined criteria or judge-based scoring, which may encourage criterion satisfaction but can provide limited signal for continual improvement.

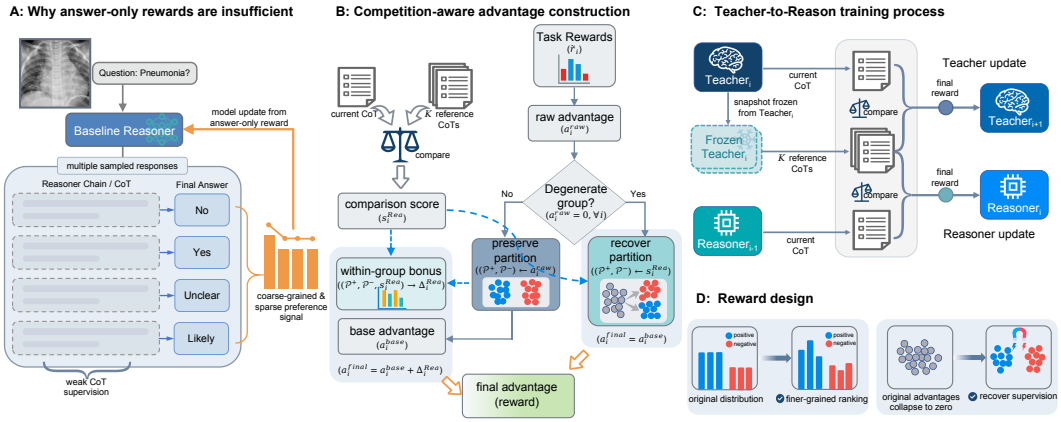


Figure 1: Overview of Teach-to-Reason (T2R). (A) Answer-only rewards provide only coarse and sparse supervision, making them insufficient for improving CoT quality in CXR VQA. (B) T2R constructs competition-aware advantages by preserving the original group partition when task rewards are informative, and recovering supervision from competition scores when the group advantage collapses to zero. (C) During training, the Teacher is iteratively improved via self-competition, and the Reasoner is optimized against progressively stronger Teacher-generated reference CoTs. (D) This yields denser and more informative training signals for CoT optimization.

We instead replace direct CoT scoring with pairwise comparison, using it as a fine-grained supervision signal that can sustain informative optimization pressure throughout training.

Self-Evolving Supervision for Reasoning. Recent work has explored how language models can improve with limited external annotation by turning their own generations, feedback, or interactions into learning signals. This includes bootstrapping from self-generated rationales, critiques, and reflections [16–18], learning from model-generated rewards or self-distilled feedback [15, 36], and constructing new tasks, preferences, or curricula through self-play, role-based competition, or environment-grounded interaction [37–44]. These works suggest that supervision for reasoning need not remain static, but can be generated, filtered, or strengthened during learning. T2R extends the self-evolving paradigm to continual CoT optimization through an iteratively improved *Teacher*. Rather than relying on direct self-refinement or self-scoring alone, T2R enables the Teacher to strengthen itself via self-competition, and uses its progressively improved CoTs to continuously guide a separate *Reasoner*. In this way, T2R uses Teacher self-evolution to drive continual improvement of the *Reasoner*, with progressively stronger Teacher-generated CoTs providing an evolving source of supervision for CoT optimization.

3 Methodology

3.1 Framework and Competition Scores

Model Roles. Our framework consists of two models: a **Reasoner** and a **Teacher**. At a high level, the Teacher provides stronger reasoning references during training, while the Reasoner is the VQA model ultimately trained for inference.

The Reasoner is the *target VQA model* in our framework. Given an image-question pair (I_i, Q_i), it generates a chain-of-thought (CoT) together with an answer:

$$(\hat{C}_i^{\text{rea}}, \hat{A}_i) \sim \pi_\theta^{\text{rea}}(I_i, Q_i). \tag{1}$$

The Teacher is used only during training. Given the image, question, reference answer, and report context, it generates higher-quality reference CoTs:

$$\hat{C}_i^{\text{tea}} \sim \pi_\phi^{\text{tea}}(I_i, Q_i, A_i^*, D_i), \tag{2}$$

where A_i^* is the reference answer and D_i denotes the radiology report or other contextual document. Rather than serving as the final VQA predictor, the Teacher provides reasoning references for the Reasoner and improves its own CoT quality through self-competition.

Comparison-Based Scoring. We use an LLM-as-a-Judge to compare the quality of two candidate reasoning chains. Given two CoTs C_1 and C_2, we define

$$f_{\text{comp}}(C_1, C_2) = \mathbf{1}[C_1 \succ C_2], \tag{3}$$

where $C_1 \succ C_2$ means that C_1 is judged better than C_2 in terms of reasoning correctness, internal consistency, answer alignment, and answer support. To mitigate position bias, we randomize the order of the two CoTs in the judge prompt. Additional implementation details of the pairwise CoT comparison are provided in Appendix A.1.

For the Reasoner, given K Teacher-generated reference CoTs $\{\hat{C}^{\text{tea}}_{i,k}\}_{k=1}^{K}$, we define the competition score as

$$s_i^{\text{rea}} = \frac{1}{K}\sum_{k=1}^{K} f_{\text{comp}}(\hat{C}_i^{\text{rea}}, \hat{C}^{\text{tea}}_{i,k}), \qquad s_i^{\text{rea}} \in [0,1]. \tag{4}$$

This score is the average win rate of the current Reasoner CoT against Teacher-generated reference CoTs. Analogously, we define the Teacher self-competition score s_i^{tea} by comparing the current Teacher output against K CoTs $\{\bar{C}^{\text{tea}}_{i,k}\}_{k=1}^{K}$ sampled from a frozen Teacher snapshot.

3.2 Reasoner Objective

Design Principle. We use a_i^{final} as the group-wise training signal for the GRPO update. When the base reward is informative, it determines the positive and negative samples, while competition scores only refine the relative ordering within each subset; otherwise, we rely on competition scores to directly construct the final training signal.

Base Reward. For a GRPO group of size G, let $\mathcal{I} = \{1, \ldots, G\}$. We first define the base task reward as

$$\tilde{r}_i^{\text{rea}} = m_i^{\text{fmt}}\big(1 + \mathbf{1}[\hat{A}_i = A_i^*]\big), \tag{5}$$

where $m_i^{\text{fmt}} \in \{0,1\}$ indicates whether the output follows the required format. We then compute the raw advantage induced by the base reward:

$$a_i^{\text{raw}} = \tilde{r}_i^{\text{rea}} - \frac{1}{G}\sum_{j=1}^{G} \tilde{r}_j^{\text{rea}}. \tag{6}$$

This naturally induces a pair of base-reward-guided subsets,

$$\mathcal{P}^+ = \{i \in \mathcal{I} : a_i^{\text{raw}} > 0\}, \qquad \mathcal{P}^- = \{i \in \mathcal{I} : a_i^{\text{raw}} < 0\}. \tag{7}$$

We distinguish two cases accordingly. If $\mathcal{P}^+ \cup \mathcal{P}^- \neq \varnothing$, then the base reward already induces a non-trivial group-wise preference, and we treat the group as non-degenerate. If $\mathcal{P}^+ = \mathcal{P}^- = \varnothing$, then all raw advantages are zero and the base reward is degenerate, in which case we instead rely on competition scores to construct the final training signal.

Non-Degenerate Case. If the base reward is non-degenerate, we directly use the subsets $\mathcal{P}^+$ and $\mathcal{P}^-$ induced by a_i^{raw}, and employ competition scores only to refine the relative ordering within each subset. Specifically, we define the subset-wise standardized patterns

$$p_i^+ = \frac{s_i^{\text{rea}} - \mu_+^{\text{rea}}}{\sigma_+^{\text{rea}}}, \quad i \in \mathcal{P}^+, \qquad p_i^- = \frac{s_i^{\text{rea}} - \mu_-^{\text{rea}}}{\sigma_-^{\text{rea}}}, \quad i \in \mathcal{P}^-, \tag{8}$$

where $\mu_+^{\text{rea}}, \sigma_+^{\text{rea}}$ and $\mu_-^{\text{rea}}, \sigma_-^{\text{rea}}$ are the mean and standard deviation of the competition scores within $\mathcal{P}^+$ and $\mathcal{P}^-$, respectively. Based on these patterns, we define the group-wise competition bonus as

$$\Delta_i^{\text{rea}} = \begin{cases} \eta_+^{\text{rea}}\, p_i^+, & i \in \mathcal{P}^+, \\ \eta_-^{\text{rea}}\, p_i^-, & i \in \mathcal{P}^-, \end{cases} \tag{9}$$

where $\eta_+^{\text{rea}} \geq 0$ and $\eta_-^{\text{rea}} \geq 0$ control the bonus strength on the positive and negative sides. To preserve the sign induced by the base reward, we define

$$\eta_+^{\text{rea}} = \alpha \min_{i \in \mathcal{P}^+,\, p_i^+ < 0} \frac{a_i^{\text{raw}}}{-p_i^+}, \qquad \eta_-^{\text{rea}} = \alpha \min_{i \in \mathcal{P}^-,\, p_i^- > 0} \frac{-a_i^{\text{raw}}}{p_i^-}, \tag{10}$$

where $\alpha \in (0,1)$ controls the safety margin. The final advantage is then

$$a_i^{\text{final}} = a_i^{\text{raw}} + \Delta_i^{\text{rea}}. \tag{11}$$

Degenerate Case. If the base reward is degenerate, it induces no non-trivial preference within the group. In this case, we construct the partition directly from the competition scores s_i^{rea}. For a candidate threshold τ, we define

$$\mathcal{P}^-(\tau) = \{i \in \mathcal{I} : s_i^{\text{rea}} \le \tau\}, \qquad \mathcal{P}^+(\tau) = \{i \in \mathcal{I} : s_i^{\text{rea}} > \tau\}. \tag{12}$$

Among all candidate thresholds, we choose the largest feasible τ satisfying

$$\sum_{i \in \mathcal{P}^-(\tau)} s_i^{\text{rea}} < \gamma \sum_{i \in \mathcal{P}^+(\tau)} s_i^{\text{rea}}, \tag{13}$$

where $\gamma > 0$ is a hyperparameter. If no feasible threshold exists, we set the final advantage of the whole group to zero.

Otherwise, after obtaining the induced subsets $\mathcal{P}^+$ and $\mathcal{P}^-$, we reconstruct the final training signal from the competition scores. We first globally center the scores:

$$p_i = s_i^{\text{rea}} - \mu^{\text{rea}}, \qquad \mu^{\text{rea}} = \frac{1}{G} \sum_{j=1}^{G} s_j^{\text{rea}}. \tag{14}$$

To make the reconstructed signal consistent with the induced partition, we then shift only the positive subset:

$$\hat{p}_i = \begin{cases} p_i + \lambda, & i \in \mathcal{P}^+, \\ p_i, & i \in \mathcal{P}^-. \end{cases} \tag{15}$$

Here λ is chosen as the minimum feasible shift that preserves the positive/negative structure implied by the partition:

$$\lambda_{\min} = \max\left(-\frac{G}{|\mathcal{P}^-|} \min_{i \in \mathcal{P}^+} p_i,\ \frac{G}{|\mathcal{P}^+|} \max_{i \in \mathcal{P}^-} p_i \right), \qquad \lambda = \max(\lambda_{\min}, 0), \tag{16}$$

with the derivation deferred to Appendix C. We then define the final advantage as

$$a_i^{\text{final}} = \hat{p}_i - \frac{1}{G} \sum_{j=1}^{G} \hat{p}_j. \tag{17}$$

3.3 Teacher Objective

The Teacher is trained to generate higher-quality CoTs that serve as stronger reasoning references for Reasoner optimization. We therefore define its reward as

$$r_i^{\text{tea}} = m_i^{\text{fmt}} \big(1 + s_i^{\text{tea}}\big), \tag{18}$$

where $m_i^{\text{fmt}} \in \{0, 1\}$ enforces output format validity and s_i^{tea} encourages improved reasoning quality through self-competition. Thus, the Teacher is optimized to maximize reasoning quality subject to valid formatting.

3.4 Training Procedure

We initialize the Teacher and the Reasoner from the same pretrained weights, denoted by $\mathcal{T}_0$ and $\mathcal{R}_0$, respectively. We first obtain $\mathcal{T}_1$ by freezing $\mathcal{T}_0$ as a reference model $\tilde{\mathcal{T}}_0$ and updating $\mathcal{T}_0$ against $\tilde{\mathcal{T}}_0$ via self-competition for K optimization steps.

Training then proceeds iteratively. At iteration i, we freeze the current Teacher $\mathcal{T}_i$ as $\tilde{\mathcal{T}}_i$ and use it as the shared reference model for both Teacher and Reasoner updates. Conditioned on the same $\tilde{\mathcal{T}}_i$, the Teacher is updated via self-competition to obtain $\mathcal{T}_{i+1}$, while the Reasoner is updated against Teacher-generated reference CoTs to obtain $\mathcal{R}_i$; both updates are performed for K optimization steps.

This yields the paired updates

$$\mathcal{T}_i \to \mathcal{T}_{i+1}, \qquad \mathcal{R}_{i-1} \to \mathcal{R}_i.$$

Repeating this procedure for T iterations produces a sequence of Reasoner models $\mathcal{R}_1, \ldots, \mathcal{R}_T$.

This training schedule is efficient because Teacher improvement is driven entirely by self-competition and does not depend on the current Reasoner. Consequently, Teacher and Reasoner updates are naturally decoupled through the same frozen Teacher reference. An overview of the training pipeline is shown in Figure 1.

4 Experiments

4.1 Experimental Setup

Models and Training Data. We use Qwen3-VL-Instruct [19] as the base model and conduct experiments at the 2B and 4B scales. Unless otherwise specified, the Teacher and the Reasoner at each scale are initialized from the same pretrained checkpoint.

Our training data is derived from RexVQA [4], a single-choice chest X-ray VQA dataset. For training efficiency, we randomly sample 67K instances. To reduce the format gap between training and evaluation, we convert 13.4K samples (approximately 20%) into open-ended form by removing the answer options, while keeping the remaining samples in their original single-choice format.

Training Details. Unless otherwise specified, we use a GRPO group size of 16 and a batch size of 128. For competition scoring, we set the number of Teacher-generated reference CoTs to $K = 10$ per sample. The safety coefficient is set to $\alpha = 0.9$, and the degenerate-case threshold is set to $\gamma = 0.3$. During training, both pairwise CoT comparison and open-ended answer verification are performed using Qwen3-4B-Instruct-2507 [45].

To instantiate the staged training framework, we partition the 67K training set into four disjoint splits: split 0 contains one-half of the data, while splits 1–3 each contain one-sixth. Let $\mathcal{T}_0$ and $\mathcal{R}_{\text{init}}$ denote the Teacher and Reasoner initialized from the same pretrained model. We first train $\mathcal{R}_{\text{init}}$ with RLVR on split 0 for one epoch to obtain $\mathcal{R}_0$. Starting from $\mathcal{T}_0$, we then train the Teacher via self-competition on splits 1–3 to obtain $\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3$. Starting from $\mathcal{R}_0$, the Reasoner is subsequently optimized stage by stage on splits 1–3 against these progressively improved Teacher models, yielding $\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3$. Additional hyperparameter settings and prompt details are provided in Appendix B.

Baselines. In addition to **Teach-to-Reason (T2R)**, we compare against three baselines. All Reasoner variants are initialized from the same $\mathcal{R}_0$ and are further trained on splits 1–3 with matched GRPO hyperparameters and training budgets, so that all methods use the same overall data budget over splits 0–3.

RLVR Only starts from $\mathcal{R}_0$ and continues RLVR on splits 1–3 without additional CoT supervision. **Frozen Teacher** starts from $\mathcal{R}_0$ and optimizes the Reasoner on splits 1–3 against Teacher-generated reference CoTs, while keeping the Teacher fixed as $\mathcal{T}_0$ throughout. **Judge Reward** replaces pairwise comparison with a binary CoT-quality signal assigned by an LLM judge according to predefined acceptability criteria. Full details of the judging criteria, prompt design, and reward formulation for **Judge Reward** are provided in Appendix A.3.

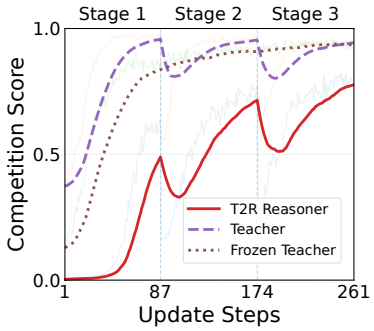
Evaluation. We evaluate our method on six CXR open-ended VQA benchmarks: Medical-CXR-VQA [46], VQA-RAD [1], SLAKE [2], MIMIC-CXR-VQA [47], Covid19_heywhale [48], and Chest_X-Ray_PA [49]. To ensure a consistent evaluation protocol, we retain only the **CXR open-ended VQA** subset from each dataset. For answer evaluation, we use Qwen3-235B-A22B [45] as the judge. Given the question, ground-truth answer, and model prediction, the judge determines whether the prediction is correct. To improve robustness, each prediction is judged five times independently, and the final correctness label is determined by majority vote. The evaluation prompt and judge hyperparameters are provided in Appendix A.2. All evaluated VQA models decode with greedy search. Unless otherwise specified, we report accuracy on all benchmarks.

4.2 Main Results

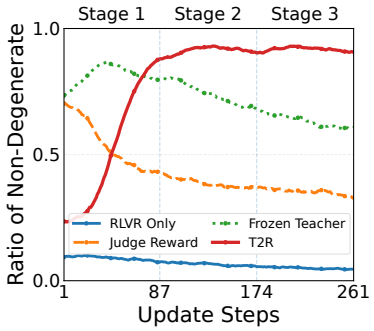
Table 1 summarizes the main benchmark results at the 2B and 4B scales. For T2R, we report three Reasoner models, denoted by **T2R-R1**, **T2R-R2**, and **T2R-R3**, corresponding to $\mathcal{R}_1$, $\mathcal{R}_2$, and $\mathcal{R}_3$, respectively. For all compared methods, we evaluate the latest checkpoint from the corresponding training run across all benchmarks. Figure 2 further shows the training dynamics of competition scores and the ratio of groups with non-zero advantages, while Figure 3 visualizes the activation of degenerate-case handling during training. Additional qualitative comparisons of generated CoTs are provided in Appendix E.

Table 1: Main results on six CXR open-ended VQA benchmarks. Best results within each model scale and dataset are shown in bold. Rows corresponding to Teach-to-Reason (T2R) are shaded in light gray and report different Reasoner stages.

Scale	Method	Datasets					
		Med-CXR	VQA-RAD	SLAKE	MIMIC-CXR	Covid19	CXR-PA
2B	Base Model	15.94	31.69	52.87	12.98	22.64	31.72
	RLVR Only	16.87	30.18	53.92	11.85	21.69	20.32
	Judge Reward	28.00	35.84	**56.02**	17.53	15.66	27.04
	Frozen Teacher	21.74	40.37	52.35	20.88	31.88	32.20
	T2R-R1	22.50	41.28	52.35	20.55	39.43	34.64
	T2R-R2	27.51	40.90	53.92	22.17	30.56	36.98
	T2R-R3	**31.73**	**43.18**	50.78	**23.87**	**40.37**	**37.57**
4B	Base Model	21.74	40.37	58.11	20.45	26.22	41.22
	RLVR Only	21.34	41.13	56.84	22.67	28.86	35.81
	Judge Reward	22.38	41.26	56.54	23.09	29.24	36.40
	Frozen Teacher	**24.41**	42.01	**64.92**	25.85	27.54	38.59
	T2R-R1	21.27	38.63	61.25	19.82	40.56	42.25
	T2R-R2	22.87	**46.96**	**64.92**	25.61	**49.43**	**45.32**
	T2R-R3	23.23	42.42	60.73	**28.04**	41.37	43.85



(a) Competition Scores



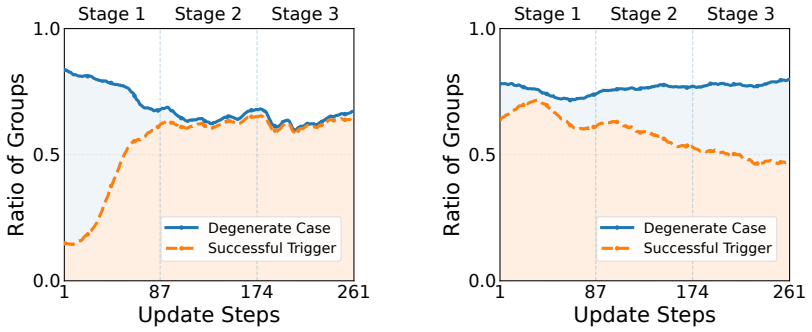
(b) Non-zero Advantage

Figure 2: Training dynamics of (a) competition scores for the T2R Reasoner, the T2R Teacher, and Frozen Teacher, and (b) the ratio of groups with non-zero advantages for RLVR Only, Judge Reward, Frozen Teacher, and T2R.

Overall Performance. T2R shows the strongest overall performance under both the 2B and 4B settings. At 2B, at least one T2R model achieves the best result on five out of six datasets, with SLAKE as the only exception, where **Judge Reward** performs best. At 4B, T2R again attains the best or tied-best result on five out of six datasets; the only exception is Medical-CXR-VQA, where **Frozen Teacher** performs best. Overall, these results indicate that comparison-guided optimization yields robust gains across diverse CXR VQA benchmarks.

Effect of Iterative Teacher Improvement. Iteratively improving the Teacher is more effective than using a frozen Teacher. As shown in Table 1, T2R achieves stronger best-case performance than **Frozen Teacher** under both the 2B and 4B settings. Figure 2(a) is consistent with this trend: both the T2R Teacher and the T2R Reasoner exhibit a stage-wise drop-and-recovery pattern, reflecting the introduction of progressively stronger comparison targets at each stage. In contrast, the Reasoner trained with **Frozen Teacher** gradually saturates, suggesting that a fixed Teacher provides increasingly limited comparison-based supervision over time.

Supervision Density During Training. T2R yields denser and more persistent supervision signals than **RLVR Only**, **Judge Reward**, and **Frozen Teacher**. As shown in Figure 2(b), **RLVR Only** maintains a consistently low ratio of groups with non-zero advantages, while **Judge Reward** and **Frozen Teacher** both decline over training. By contrast, T2R rises quickly and remains at the highest level in later stages, indicating that iterative Teacher improvement sustains a substantially larger



(a) T2R (b) Frozen Teacher

Figure 3: Degenerate-case statistics during training. (a) Ratios of degenerate groups and successfully activated degenerate-case handling groups during T2R training. (b) The same statistics for Frozen Teacher.

Table 2: Ablation on the threshold hyperparameter γ for degenerate-group partitioning.

γ	Med-CXR	VQA-RAD	SLAKE	MIMIC-CXR	Covid-19	CXR-PA
0.1	**28.11**	41.00	51.96	22.16	34.21	**37.26**
0.3	27.24	**41.78**	**52.35**	**22.19**	**36.78**	36.39
0.6	26.73	36.59	51.48	21.80	35.33	36.24
0.9	27.01	38.61	50.25	21.64	33.86	35.16

Table 3: Effect of Teacher strength in the Frozen Teacher setting.

Teacher	Med-CXR	VQA-RAD	SLAKE	MIMIC-CXR	Covid-19	CXR-PA
$\mathcal{T}_0$	21.74	**40.37**	52.35	**20.88**	31.88	32.20
$\mathcal{T}_1$	**23.78**	37.50	51.30	19.70	**34.90**	**34.88**
$\mathcal{T}_2$	18.03	32.19	**55.49**	13.57	18.30	25.14
$\mathcal{T}_3$	17.52	33.33	51.30	13.31	22.64	24.26

fraction of trainable groups throughout optimization. This is consistent with the stronger benchmark performance of T2R in Table 1.

Impact of Degenerate-Case Handling. Figure 3 shows that degenerate-case handling is not merely activated occasionally, but remains highly engaged throughout T2R training. Under T2R, the ratio of successfully triggered groups rises quickly and stays close to the overall ratio of degenerate groups, indicating that the proposed design effectively restores supervision when the original group reward collapses. In contrast, under **Frozen Teacher**, the trigger ratio gradually decreases over time despite the continued presence of degenerate groups. This pattern highlights the importance of iterative Teacher improvement for sustaining supervision in degenerate groups.

4.3 Ablation Study

Effect of γ. We study the effect of the threshold hyperparameter γ in degenerate-case partitioning. Table 2 reports results under the 2B setting, where each entry is the mean accuracy of the three Reasoner models $\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3$ on the corresponding benchmark. Overall, $\gamma = 0.3$ achieves the best trade-off, yielding the best result on four out of six datasets.

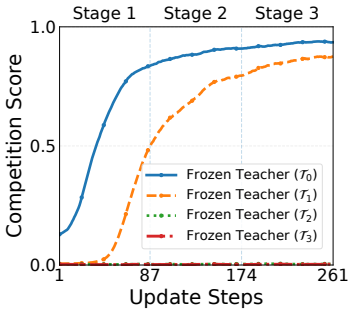


Figure 4: Competition score dynamics under different fixed Teachers in Frozen Teacher setting.

The role of γ is to control how aggressively degenerate groups are partitioned using competition scores. Larger values make the partition more permissive, so some samples that are not clearly inferior may still be assigned to the negative subset, reducing the reliability of the reconstructed training signal. Smaller values make the partition more conservative, so some potentially useful degenerate groups may fail to trigger the reconstruction mechanism. Consistent with this intuition, overly large values ($\gamma = 0.6, 0.9$) perform worse overall, while an overly small value such as $\gamma = 0.1$ is slightly better on a few datasets but remains less favorable. We therefore adopt $\gamma = 0.3$ as the default setting.

Table 4: Component ablations under the 2B setting. *Degenerate Handling* keeps only degenerate-case handling, *Non-Degenerate Bonus* keeps only the non-degenerate bonus, and *Direct Competition Reward* directly injects competition scores into the reward.

Method	Med-CXR	VQA-RAD	SLAKE	MIMIC-CXR	Covid19	CXR-PA
Base Model	15.94	31.69	<u>52.87</u>	12.98	22.64	31.72
RLVR Only	16.87	30.18	**53.92**	11.85	21.69	20.32
Degenerate Handling	**28.50**	<u>40.62</u>	51.30	<u>21.20</u>	**45.71**	<u>35.07</u>
Non-Degenerate Bonus	20.94	31.69	**53.92**	12.52	20.93	25.58
Direct Competition Reward	26.55	38.74	51.04	21.14	33.77	34.63
T2R (ours)	<u>27.24</u>	**41.78**	52.35	**22.19**	<u>36.78</u>	**36.39**

Table 5: Effect of the comparison judge on 4B T2R. All other training and evaluation settings are kept unchanged.

Judge	Stage	Med-CXR	VQA-RAD	SLAKE	MIMIC-CXR	Covid19	CXR-PA	Avg.
Qwen3-4B-Instruct-2507	T2R-R1	21.27	38.63	61.25	19.82	40.56	42.25	37.30
	T2R-R2	22.87	<u>46.96</u>	<u>64.92</u>	25.61	49.43	45.32	42.52
	T2R-R3	23.23	42.42	60.73	<u>28.04</u>	41.37	43.85	39.94
Qwen3-8B	T2R-R1	24.26	41.78	63.23	24.23	48.49	42.69	40.78
	T2R-R2	<u>31.06</u>	43.64	**65.45**	26.59	<u>50.77</u>	<u>47.88</u>	<u>44.23</u>
	T2R-R3	**33.47**	**48.52**	61.83	**29.77**	**54.52**	**50.14**	**46.38**

Component Ablation. We ablate the main competition-based components of T2R in Table 4. **Degenerate Handling** retains only the competition-guided partition and supervision recovery in degenerate groups. **Non-Degenerate Bonus** retains only the sign-preserving bonus in non-degenerate groups. **Direct Competition Reward** removes the case-wise design and instead directly injects the competition score into the training reward:

$$\tilde{r}_i^{\mathrm{rea}} = m_i^{\mathrm{fmt}}\big(1 + s_i^{\mathrm{rea}} + \mathbf{1}[\hat{A}_i = A_i^*]\big). \tag{19}$$

As shown in Table 4, **Degenerate Handling** yields larger gains than **Non-Degenerate Bonus** on most datasets, suggesting that a major part of T2R's improvement comes from recovering supervision in degenerate groups. By contrast, **Non-Degenerate Bonus** still improves over **RLVR Only** on most datasets, indicating that within-group refinement is helpful but not the dominant source of gain.

Meanwhile, **Direct Competition Reward**, although stronger than **Base Model** and **RLVR Only**, remains weaker than both **Degenerate Handling** and the full **T2R**. This suggests that competition scores are most effective when incorporated into the reward in a controlled and principled way, rather than being directly injected into the original reward.

Importance of a Comparable Teacher. We use the **Frozen Teacher** setting to study how Teacher strength affects Reasoner training, by fixing the Teacher as $\mathcal{T}_0$, $\mathcal{T}_1$, $\mathcal{T}_2$, or $\mathcal{T}_3$. Figure 4 shows that the Reasoner's competition score increases over training when the fixed Teacher is $\mathcal{T}_0$ or $\mathcal{T}_1$, but remains near zero when it is $\mathcal{T}_2$ or $\mathcal{T}_3$. Consistently, Table 3 shows that $\mathcal{T}_0$ and $\mathcal{T}_1$ yield much better results. These observations suggest that effective comparison-based supervision requires a Teacher of comparable difficulty: if the Teacher is too strong, the comparison signal becomes overly sparse, leaving little room for the Reasoner to improve. This also motivates the stage-wise design of T2R, where the Teacher is updated after each stage so that it remains comparable to the current Reasoner.

4.4 Effect of the Comparison Judge

We further examine how the choice of comparison judge affects T2R training. For the 4B Reasoner, we replace `Qwen3-4B-Instruct-2507` with `Qwen3-8B` as the pairwise comparison judge and repeat the full three-stage training. All other settings remain unchanged, including the training data, model initialization, stage partitions, optimization settings, Teacher reference-CoT sampling, training-time answer verifier, and final evaluation protocol.

The stage-wise results in Table 1 exhibit different trends at the two model scales. At 2B, performance generally improves through R3, which achieves the strongest overall results. At 4B, however,

Table 6: CoT quality analysis via SFT distillation under the 2B setting.

Method	Med-CXR	VQA-RAD	SLAKE	MIMIC-CXR	Covid19	CXR-PA
RLVR Only	13.79	28.03	51.30	14.28	23.96	22.07
Judge Reward	21.40	33.71	51.83	16.94	27.73	26.60
Frozen Teacher	<u>23.71</u>	<u>36.36</u>	**56.54**	<u>20.99</u>	<u>27.90</u>	<u>32.98</u>
T2R-R3	**26.44**	**39.77**	<u>56.02</u>	**21.28**	**30.75**	**33.04**

performance declines from R2 to R3 on four of the six benchmarks. As shown in Table 5, replacing the original comparison judge with Qwen3-8B improves performance at both R2 and R3. In particular, T2R-R3 outperforms its counterpart trained with the original judge on all six benchmarks, with an average gain of 6.44 percentage points. With the stronger judge, R3 also outperforms R2 on five of the six benchmarks, in contrast to the R2-to-R3 performance decline observed with the original judge.

These results suggest that the original comparison judge may limit later-stage T2R optimization in the 4B setting. As the Reasoner improves, distinguishing between candidate CoTs becomes more challenging, particularly for a smaller comparison judge. This highlights the importance of using a stronger comparison judge to sustain performance improvements throughout T2R training. Further analyses of the comparison judge are provided in Appendices D.4 and D.5.

4.5 CoT Quality Analysis via SFT Distillation

We further analyze the quality of the generated *CoT + answer* rollouts through an SFT distillation protocol. Specifically, we randomly sample 20K inputs from the 67K training set and use **RLVR Only**, **Judge Reward**, **Frozen Teacher**, and **T2R-R3** to generate one rollout per input using greedy decoding. Each rollout set is then used to supervise the same Qwen3-VL-2B-Instruct student model for one epoch under identical initialization and identical SFT hyperparameters. The resulting students are evaluated on the same six CXR open-ended VQA benchmarks.

Although this protocol does not measure CoT quality in complete isolation, it provides a more controlled proxy than directly asking an LLM to compare CoTs, which may be affected by position bias, length bias, and judge preference. If the rollouts from one method consistently yield a stronger student under the same SFT budget, this provides evidence that its generated supervision is of higher practical quality.

Table 6 shows a clear progression across the four rollout sources. **RLVR Only** yields the weakest student across the six benchmarks, suggesting that improving answers alone is insufficient for producing high-quality reasoning traces and that explicit CoT supervision is necessary. Both **Frozen Teacher** and **T2R-R3** then consistently outperform **Judge Reward**, indicating that comparison-based supervision is more effective than direct judge-based reward for improving CoT quality. Finally, among the comparison-based methods, **T2R-R3** further outperforms **Frozen Teacher** on most benchmarks, suggesting that iterative Teacher refinement further improves the quality of the generated CoTs, as reflected by the stronger student distilled from **T2R-R3** rollouts.

5 Conclusion

We presented **Teach-to-Reason (T2R)**, a training framework for CXR VQA in which an iteratively self-improving Teacher provides progressively stronger references for a comparison-guided Reasoner, while the original reward-induced partition is preserved whenever it is informative and supervision is restored from competition scores when the original signal degenerates.

More broadly, T2R suggests a new training paradigm in which *comparison*, rather than direct score assignment, serves as the reward signal for reinforcement learning. In this work, comparison is used only to judge which CoT is better overall, but in future work, the same idea could be extended much further by defining multiple finer-grained comparisons, each tied to a specific rule, preference, or task requirement. Moreover, future work could explore extending this paradigm beyond CoT optimization to other model outputs, such as solution strategies, plans, code, or other structured outputs, pointing toward more controllable RL-based training through customizable comparison signals.

Acknowledgements

This work was supported by the AI Research Center of Midea Group. The authors declare no competing interests beyond the affiliations and funding disclosed in this paper.

References

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

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

[3] Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei JI, Eric Chang, Tackeun Kim, et al. Mimic-ext-mimic-cxr-vqa: a complex, diverse, and large-scale visual question answering dataset for chest x-ray images. *PhysioNet*, 2024.

[4] Ankit Pal, Jung-Oh Lee, Xiaoman Zhang, Malaikannan Sankarasubbu, Seunghyeon Roh, Won Jung Kim, Meesun Lee, and Pranav Rajpurkar. Rexvqa: A large-scale visual question answering benchmark for generalist chest x-ray understanding. In *Biocomputing 2026: Proceedings of the Pacific Symposium*, pages 251–264. World Scientific, 2025.

[5] Xiaotang Gai, Chenyi Zhou, Jiaxiang Liu, Yang Feng, Jian Wu, and Zuozhu Liu. Medthink: A rationale-guided framework for explaining medical visual question answering. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7438–7450, 2025.

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

[7] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[8] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

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

[10] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

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

[12] Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. *arXiv preprint arXiv:2507.17746*, 2025.

[13] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *The Innovation*, 2024.

[14] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

[15] Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E Weston. Self-rewarding language models. In *Forty-first International Conference on Machine Learning*, 2024.

[16] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

[17] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

[18] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

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

[20] Asma Ben Abacha, Sadid A Hasan, Vivek V Datla, Dina Demner-Fushman, and Henning Müller. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes*. 9-12 September 2019, 2019.

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

[22] Yash Khare, Viraj Bagal, Minesh Mathew, Adithi Devi, U Deva Priyakumar, and CV Jawahar. Mmbert: Multimodal bert pretraining for improved medical vqa. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1033–1036. IEEE, 2021.

[23] Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. Multi-modal masked autoencoders for medical vision-and-language pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 679–689. Springer, 2022.

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

[25] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023.

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

[27] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

[28] Zhihong Lin, Donghao Zhang, Qingyi Tao, Danli Shi, Gholamreza Haffari, Qi Wu, Mingguang He, and Zongyuan Ge. Medical visual question answering: A survey. *Artificial Intelligence in Medicine*, 143: 102611, 2023.

[29] Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

[30] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The twelfth international conference on learning representations*, 2023.

[31] Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, et al. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

[32] Zenan Huang, Yihong Zhuang, Guoshan Lu, Zeyu Qin, Haokai Xu, Tianyu Zhao, Ru Peng, Jiaqi Hu, Zhanming Shen, Xiaomeng Hu, et al. Reinforcement learning with rubric anchors. *arXiv preprint arXiv:2508.12790*, 2025.

[33] Rulin Shao, Akari Asai, Shannon Zejiang Shen, Hamish Ivison, Varsha Kishore, Jingming Zhuo, Xinran Zhao, Molly Park, Samuel G Finlayson, David Sontag, et al. Dr tulu: Reinforcement learning with evolving rubrics for deep research. *arXiv preprint arXiv:2511.19399*, 2025.

[34] Junkai Zhang, Zihao Wang, Lin Gui, Swarnashree Mysore Sathyendra, Jaehwan Jeong, Victor Veitch, Wei Wang, Yunzhong He, Bing Liu, and Lifeng Jin. Chasing the tail: Effective rubric-based reward modeling for large language model post-training. *arXiv preprint arXiv:2509.21500*, 2025.

[35] Yang Zhou, Sunzhu Li, Shunyu Liu, Wenkai Fang, Kongcheng Zhang, Jiale Zhao, Jingwen Yang, Yihe Zhou, Jianwei Lv, Tongya Zheng, et al. Breaking the exploration bottleneck: Rubric-scaffolded reinforcement learning for general llm reasoning. *arXiv preprint arXiv:2508.16949*, 2025.

[36] Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, et al. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026.

[37] Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*, 2024.

[38] Wenkai Fang, Shunyu Liu, Yang Zhou, Kongcheng Zhang, Tongya Zheng, Kaixuan Chen, Mingli Song, and Dacheng Tao. Serl: Self-play reinforcement learning for large language models with limited data. *arXiv preprint arXiv:2505.20347*, 2025.

[39] Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. Absolute zero: Reinforced self-play reasoning with zero data. *arXiv preprint arXiv:2505.03335*, 2025.

[40] Bo Liu, Chuanyang Jin, Seungone Kim, Weizhe Yuan, Wenting Zhao, Ilia Kulikov, Xian Li, Sainbayar Sukhbaatar, Jack Lanchantin, and Jason Weston. Spice: Self-play in corpus environments improves reasoning. *arXiv preprint arXiv:2510.24684*, 2025.

[41] Hongliang Lu, Yuhang Wen, Pengyu Cheng, Ruijin Ding, Jiaqi Guo, Haotian Xu, Chutian Wang, Haonan Chen, Xiaoxi Jiang, and Guanjun Jiang. Search self-play: Pushing the frontier of agent capability without supervision. *arXiv preprint arXiv:2510.18821*, 2025.

[42] Jingyu Zhang, Haozhu Wang, Eric Michael Smith, Sid Wang, Amr Sharaf, Mahesh Pasupuleti, Benjamin Van Durme, Daniel Khashabi, Jason Weston, and Hongyuan Zhan. The alignment waltz: Jointly training agents to collaborate for safety. *arXiv preprint arXiv:2510.08240*, 2025.

[43] Yicheng He, Chengsong Huang, Zongxia Li, Jiaxin Huang, and Yonghui Yang. Visplay: Self-evolving vision-language models from images. *arXiv preprint arXiv:2511.15661*, 2025.

[44] Huan ang Gao, Jiayi Geng, Wenyue Hua, Mengkang Hu, Xinzhe Juan, Hongzhang Liu, Shilong Liu, Jiahao Qiu, Xuan Qi, Yiran Wu, et al. A survey of self-evolving agents: On path to artificial super intelligence, 2025. URL `https://arxiv.org/abs/2507.21046`.

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

[46] Xinyue Hu, Lin Gu, Kazuma Kobayashi, liangchen liu, Mengliang Zhang, Tatsuya Harada, Ronald Summers, and Yingying Zhu. Medical-CXR-VQA dataset: A Large-Scale LLM-Enhanced Medical Dataset for Visual Question Answering on Chest X-Ray Images. *PhysioNet*, January 2025. doi: 10.13026/1pm5-hy02. URL `https://doi.org/10.13026/1pm5-hy02`. Version 1.0.0.

[47] Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei Ji, Eric Chang, Tackeun Kim, et al. Ehrxqa: A multi-modal question answering dataset for electronic health records with chest x-ray images. *Advances in Neural Information Processing Systems*, 36:3867–3880, 2023.

[48] Muhammad EH Chowdhury, Tawsifur Rahman, Amith Khandakar, Rashid Mazhar, Muhammad Abdul Kadir, Zaid Bin Mahbub, Khandakar Reajul Islam, Muhammad Salman Khan, Atif Iqbal, Nasser Al Emadi, et al. Can ai help in screening viral and covid-19 pneumonia? *Ieee Access*, 8:132665–132676, 2020.

[49] A Asraf and Z Islam. Covid19 pneumonia and normal chest x-ray pa dataset. mendeley data v1 (2021), 2021.

[50] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL `https://arxiv.org/abs/2504.10479`.

[51] M2 Team, Chengfeng Dou, Chong Liu, Fan Yang, Fei Li, Jiyuan Jia, Mingyang Chen, Qiang Ju, Shuai Wang, Shunya Dang, Tianpeng Li, Xiangrong Zeng, and et al. Yijie Zhou. Baichuan-m2: Scaling medical capability with large verifier system, 2025. URL `https://arxiv.org/abs/2509.02208`.

A LLM-as-a-Judge Details

Unless otherwise specified, we use `Qwen3-4B-Instruct-2507` as the judge model and deploy it with `vLLM`, using `max_model_len=32768`, `gpu_mem_util=0.9`, `max_tokens=2048`, `temperature=0.7`, and `top_p=0.9`. Across all judging tasks, the model is required to return a valid XML response with a task-specific label in the `<result>` field. For binary yes/no judging tasks, if the XML output is malformed or the `<result>` field is invalid, we apply a unified fallback rule: if the decoded response contains the substring `>no<` in a case-insensitive manner, we assign `no`; otherwise, we assign `yes`.

A.1 CoT Comparison for Competition Scoring

For pairwise CoT comparison during training, the judge receives a chest X-ray report, an exam question, the official correct answer, and two candidate reasoning traces. It is asked to determine which reasoning trace is overall better, based on content quality rather than length. The comparison criteria include reasoning correctness, internal consistency, answer alignment, answer support, relevance, restraint, and clarity. The comparison prompt is shown in Figure 5.

To mitigate position bias, the two candidate CoTs are randomly assigned to the prompt slots `{reason_a}` and `{reason_b}`. The judge must return exactly one winner, either `A` or `B`; tie outputs are not allowed. After decoding, the predicted label is mapped back to the original ordering of the compared CoTs before computing the comparison outcome.

For this task, malformed XML outputs or invalid `<result>` labels are discarded rather than repaired. After filtering invalid comparisons, we compute the competition score using the retained valid outputs. The Reasoner competition score is

$$s_i^{\mathrm{rea}} = \frac{1}{K}\sum_{k=1}^{K} f_{\mathrm{comp}}(\hat{C}_i^{\mathrm{rea}}, \hat{C}_{i,k}^{\mathrm{tea}}), \qquad s_i^{\mathrm{rea}} \in [0,1], \tag{20}$$

and the Teacher self-competition score is defined analogously as

$$s_i^{\mathrm{tea}} = \frac{1}{K}\sum_{k=1}^{K} f_{\mathrm{comp}}(\hat{C}_i^{\mathrm{tea}}, \bar{C}_{i,k}^{\mathrm{tea}}), \qquad s_i^{\mathrm{tea}} \in [0,1]. \tag{21}$$

A.2 Open-Ended Answer Judging

For open-ended answer grading, the judge receives a radiology report, an open-ended question, a reference answer, and a predicted answer, and determines whether the prediction should be considered correct. The prompt instructs the judge to ground its decision primarily in the radiology report, while using the reference answer to identify the core medical elements required by the question. The grading prompt is shown in Figure 6.

Training-time setting. During training, we use the default judge model and configuration described above. Each sample is judged once, and the judge is required to return a binary decision with `<result>` equal to either `yes` or `no`. When the XML output is malformed or the label is invalid, we apply the unified fallback rule so that each training sample yields a binary supervision signal.

Final-evaluation setting. At final evaluation time, we replace the default judge with a stronger model, `Qwen3-235B-A22B`, while keeping the same prompt template and per-run decoding configuration. Each sample is judged five times independently, and the final label is determined by majority vote over the five binary decisions. The same fallback rule is applied to each run before aggregation. This repeated-judging strategy stabilizes the final LLM-as-a-Judge decision and reduces evaluation variance.

A.3 Reasoning Quality Assessment for Auxiliary Reward

In the comparative experiments, we additionally consider a **Judge Reward** baseline, in which an LLM evaluates whether a single reasoning trace is acceptable and converts this judgment into an auxiliary

Prompt template used for pairwise CoT comparison

```
You will be given a chest X-ray exam question, a corresponding radiology report, the official correct answer, and two different reasoning traces (labeled A and B). Both reasoning traces attempt to explain how one could arrive at the official answer based on the chest X-ray and the question.

Your task: after fully understanding the case and the question, compare reasoning A and reasoning B, decide which one is overall better, and briefly explain your judgment.

【Available information】

* Radiology report (reflects the true findings and diagnostic thinking for this case):
  {report}

* Exam question:
  {question}

* Official correct answer:
  {answer}

* Reasoning A:
  {reason_a}

* Reasoning B:
  {reason_b}

【Overall evaluation principle】
Your judgment of which reasoning is "better" must be based on content quality, not on length, word count, number of sentences, or superficial level of detail.

* Do not judge a reasoning trace as better simply because it is longer, more verbose, or appears more detailed;
* Do not judge a reasoning trace as worse simply because it is shorter;
* A shorter reasoning trace can be better if it is more accurate, more consistent with the imaging findings, and more logically focused;
* A longer reasoning trace should not be rewarded if it contains redundancy, repetition, irrelevant discussion, unsupported speculation, or incorrect reasoning;
* You must compare A and B based only on the quality of their content: correctness, relevance, reasoning quality, restraint, and clarity. Length itself must not be treated as an advantage or disadvantage.

【Evaluation criteria】
When comparing A and B, you should consider the following aspects. It is allowed that both have flaws; you only need to choose the relatively better one.

1. Correctness and consistency

   * Does the reasoning logically support the official answer {answer}, instead of drifting toward a different answer?
   * Is it broadly consistent with the radiology report in terms of imaging findings and diagnostic direction?
   * Does it avoid obvious medical errors or reasoning that conflicts with basic chest imaging knowledge?

2. Use of evidence and reasoning chain

   * Does it clearly build on plausible key findings that one could see on the chest X-ray, rather than just restating the answer or giving vague comments?
   * Does it show a reasonable reasoning flow, for example:
     "inspect the image → describe key abnormalities/locations/extent → relate to the question → rule out less likely options → reach the conclusion"?
   * Is the reasoning coherent and causal, rather than fragmented and disjointed?

3. Information use and restraint

   * Does it focus on the findings and information that are actually relevant to this specific question, instead of introducing large amounts of unrelated content?
   * Does it avoid clearly inventing findings or test results that are not supported by the image or the case?
   * If it implicitly aligns with ideas present in the radiology report, does it remain compatible with the spirit of the report, and does it sound like it could come from reading the film rather than obviously "quoting the report"?
   * Does it avoid unnecessary elaboration added merely to sound more complete? Relevant and useful content matters more than more content.

4. Clarity of expression

   * Is the language clear and easy to understand?
   * Does it clearly explain why the official answer makes sense, rather than merely restating the conclusion?
   * Is it concise but effective, without losing focus through excessive expansion?

【Special instructions for comparison】

* This is a quality comparison, not a length comparison;
* Length should be treated only as a surface characteristic, not as a scoring factor by itself;
* A longer reasoning trace only deserves credit if the additional content is genuinely correct, relevant, necessary, and helpful to the reasoning process;
* If the extra length mainly consists of repetition, vague expansion, irrelevant content, or unsupported details, that should count against it rather than in its favor;
* If A and B are close in quality, prefer the one that is more accurate, more focused, and better grounded in key findings, rather than defaulting to the longer one.

【Output requirements】

1. Internally compare reasoning A and B using the above criteria, and decide which one is overall better.
2. Then output your decision and justification in XML format, with root tag <response> and two child tags:

   * <reason>: briefly explain how you compared A and B, in which aspects A is better or B is better, and why you finally chose one over the other.
   * <result>: write only a single capital letter, "A" or "B", indicating which reasoning you judge to be better overall.

Important:

* Do NOT output anything outside the XML structure;
* Inside <result>, output only a single capital letter A or B, without quotes or any extra text;
* In <reason>, do not use "longer" or "shorter" itself as the main justification unless you explicitly connect it to an actual content-level strength or weakness.

Please output in the following format:

```xml
<response>
    <reason>(briefly describe your comparison and justification, e.g., which reasoning is more consistent with the report and the official answer, which has a more coherent chain of thought, whether there are obvious medical mistakes, etc.)</reason>
    <result>(write A or B here, corresponding to the reasoning you consider better)</result>
</response>
```

Your output:
```

Figure 5: Prompt template used for pairwise CoT comparison in the LLM-as-a-Judge module. The two candidate reasoning traces are randomly assigned to slots A and B to reduce position bias.

Prompt template for grading of open-ended answers

```
You are now acting as a grading examiner for chest radiology open-ended exam questions.

Your task:
Given a 【chest X-ray report】, an 【open-ended question】, a 【reference (ground truth) answer】 and a 【student answer】,
decide whether the student's answer should be considered correct, and briefly explain your reasoning.

【Information】

1. Chest X-ray report (Report)
  - This is the factual basis for the case.
  - If the reference answer or the student answer conflicts with the report, the report has the highest priority for judging medical correctness.

2. Reference answer (Ground Truth)
  - Provided by the instructor; it reflects the key information that the question is intended to test.
  - You should extract the core medical elements from it (e.g. abnormal findings, diagnosis, location, cause, extent, severity, management advice).

3. Student answer (Pred)
  - The wording may differ from the reference answer.
  - As long as the medical meaning is equivalent or very close, and it does not contradict the report, it can be graded as correct.

【Grading principles】

You must decide a binary result (yes/no) based on the following:

1. Focus on what the question is asking for
  - First, understand what type of information the question explicitly asks:
    - e.g. "most likely diagnosis", "main abnormal finding", "location of the lesion", "possible cause", "severity", "management step", etc.
  - The reference answer shows what the instructor really wants the student to provide.

2. Identify the core elements in the reference answer
  - The reference answer may contain one or several key points:
    - If the question asks for "main diagnosis / main abnormality / most important change", the student answer should at least cover that main core element; minor omissions may be acceptable.
    - If the question explicitly asks for "all major abnormalities / list all …", then missing any important element should usually be treated as not fully correct (result = "no").

3. When to mark the student as correct (result = "yes")
  - The student's answer matches the core meaning of the reference answer:
    - Synonyms, paraphrases, and equivalent medical terminology are acceptable;
    - Shorter wording is acceptable if it still captures the key medical content.
  - The student answer must NOT:
    - Present a diagnosis/finding/location/cause that clearly contradicts the report or the reference answer;
    - Be overly vague (e.g. "something is wrong", "there is an abnormality") while the question expects a specific diagnosis or finding;
    - Make clear mistakes on crucial aspects (e.g. wrong side, wrong structure, wrong type of pathology, wrong severity grade).

4. When to mark the student as incorrect (result = "no")
  - The student answer misses the key information required by the question;
  - The main content of the student answer conflicts with the reference answer and/or the report;
  - The answer is too general or ambiguous and does not demonstrate actual understanding of the required specific point;
  - The answer contains serious medical errors, even if some parts sound similar to the reference.

5. Leniency vs strictness
  - Do not require exact wording; judge based on medical meaning.
  - Be tolerant of minor phrasing differences that do not affect correctness.
  - Be strict regarding the main diagnostic direction, key location, main abnormality type, key cause, and other crucial elements.

【Output format】

1. First complete your reasoning internally; do NOT output your intermediate thoughts.
2. Then output exactly one XML code block, wrapped in ```xml, with the following structure:

```xml
<response>
   <reason>Use 2–5 sentences in English to briefly explain why you judged the student answer as correct or incorrect. Mention what the question asks for, what the core elements of the reference answer are, and whether the student answer matches them or contains clear errors/omissions.</reason>
   <result>yes or no (lowercase only)</result>
</response>
````

3. Do NOT output anything outside this XML code block. No extra explanations, no additional code blocks.

【Chest X-ray report】
{report}

【Open-ended question】
{question}

【Reference (ground truth) answer】
{ground_truth}

【Student answer】
{pred}

Based on the above information and grading principles, decide whether the student's answer should be considered correct,
and return your XML response strictly following the required format.

Your Prediction:
```

Figure 6: Prompt template used for LLM-as-a-Judge grading of open-ended answers. The judge determines whether the predicted answer is medically correct given the radiology report, the question, and the reference answer.

Prompt template used for reasoning quality assessment

You will be given a chest X-ray exam question, a corresponding radiology report, the official correct answer, and one reasoning trace (chain of thought) for solving this question.

Your task: using all this information, decide whether this reasoning trace is “acceptable” or “not acceptable”, and explain your judgment. Here, “acceptable” means that the reasoning can reasonably serve as a reliable, instructive solution process for this question.

【Available information】

- Radiology report (can be regarded as the imaging “ground truth” for this case):
{report}

- Exam question:
{question}

- Official correct answer:
{answer}

- Reasoning to be evaluated:
{reason}

【Evaluation criteria – key dimensions】

1. Consistency with the official answer
  - Does the reasoning ultimately support the given official answer {answer} (at least not contradict it in meaning)?
  - Does it clearly explain why this answer is reasonable, rather than implicitly suggesting that some other answer would be more appropriate?

2. Medical consistency with the case/report
  - Do the imaging findings, abnormalities, diagnostic tendencies, etc. mentioned in the reasoning broadly match the radiology report in terms of direction and key facts?
  - Does it avoid conclusions that are clearly opposite to the report or strongly violate basic chest imaging knowledge (e.g., treating obviously normal structures as severe pathology, or claiming an abnormality is completely absent when the report clearly states it is present)?

3. Completeness and plausibility of the reasoning chain
  - Does it present a genuine reasoning process, rather than just a bare conclusion or a single sentence explanation?
  - Does it cover most of the following elements:
    “inspect the image → describe key findings → analyze in light of the question → rule out less likely possibilities → arrive at a conclusion that matches the official answer”?
  - Is the reasoning coherent and causal, instead of fragmented and logically unclear?

4. Proper use of information
  - Does it mainly rely on findings that could reasonably be observed on the chest X-ray and on the information in the question, rather than inventing non-existent tests or imaging features?
  - Does it avoid obviously copying phrases from the radiology report, or explicitly revealing that it saw the report/answer (e.g., “the report states…”, “the correct answer is…”)?
  - Even if it is aligned with the spirit of the report, it should read like reasoning based on image interpretation, not like quoting the report as the main justification.

5. Clarity of explanation
  - Is the language clear enough that someone with basic medical background can understand what the reasoning is saying?
  - Does it stay focused on what this particular question is asking, without long digressions into irrelevant content?

6. Language quality and readability
  - Is the text reasonably fluent, with basically correct grammar, and without so many typos, nonsense words, or scrambled word order that it becomes hard to understand?
  - Check whether the main language of the reasoning is English:
    - It may include some non-English technical terms or abbreviations, but the majority of the text should be readable English.
    - If most of the reasoning is in another language, or consists largely of gibberish/garbled characters, it should be judged as not acceptable.

【Decision rule (you make the final call)】

You should integrate all the above dimensions and decide whether the reasoning is “acceptable”:

- If overall the reasoning is medically sound, logically reasonable, supports the official answer, does not seriously violate the constraints (e.g., does not blatantly quote the report or admit to knowing the official answer), and the language quality is basically acceptable, you may judge it as **yes (acceptable)**.
- If there are serious problems—such as strong conflict with the report and basic medical knowledge, failure to support the official answer, extremely chaotic or incoherent logic, explicit signs of “I saw the report/answer and I’m explaining backwards”, or very poor language/mostly non-English so that it is hard to read—then you should judge it as **no (not acceptable)**.
- Minor imperfections or slightly suboptimal wording are allowed; the key question is whether it can reasonably function as a trustworthy “standard reasoning process” for this question.

【Output requirements】

1. First, complete your internal evaluation, then provide your judgment and justification in XML format.
2. Use <response> as the root tag, with two child tags:
  - <reason>: briefly explain, in natural English, why you consider the reasoning acceptable or not acceptable. You may mention both strengths and major flaws.
  - <result>: write yes or no, indicating your final judgment on the reasoning trace.

Important:
- Do NOT output anything outside the XML structure.
- Inside <result>, you must output only the lowercase word yes or no, without quotes or any other characters.

Please output in the following format:

```xml
<response>
  <reason>(Explain your basis and reasoning here, e.g., whether it matches the official answer and the report, whether the reasoning is complete and coherent, whether there are obvious medical errors or signs of having seen the report/answer, and whether the language is readable and mainly in English.)</reason>
  <result>(write yes or no here)</result>
</response>
```

Your Predict:

Figure 7: Prompt template used for reasoning quality assessment in auxiliary reward construction. The judge determines whether a reasoning trace can serve as an acceptable and reliable solution process for the given chest X-ray question.

reward signal. Specifically, given a radiology report, an exam question, the official correct answer, and one reasoning trace, the judge determines whether the reasoning can serve as a medically sound, logically coherent, and pedagogically reliable solution process for the question. The corresponding prompt is shown in Figure 7.

The judge returns a binary XML decision with `<result>` equal to either `yes` or `no`. We map this decision to

$$s_i^{\text{judge}} \in \{0, 1\}, \tag{22}$$

where `yes` corresponds to $s_i^{\text{judge}} = 1$ and `no` corresponds to $s_i^{\text{judge}} = 0$. This binary signal is then used to construct the reward in the judge-reward baseline:

$$\tilde{r}_i^{\text{rea}} = m_i^{\text{fmt}}\big(1 + s_i^{\text{judge}} + \mathbf{1}[\hat{A}_i = A_i^*]\big). \tag{23}$$

The prompt asks the judge to evaluate the reasoning in terms of answer consistency, medical consistency with the report, coherence and plausibility of the reasoning process, proper use of available information, clarity of explanation, and overall readability. In essence, a reasoning trace is accepted only when it is broadly consistent with both the report and the official answer, follows a coherent and medically plausible reasoning process, avoids obvious factual conflicts or unsupported speculation, and remains readable.

As in training-time open-ended answer judging, malformed XML outputs or invalid `<result>` labels are handled by the same fallback rule described above. Each reasoning trace is judged once in this setting.

B Training Details

B.1 Training Setup

We optimize both the Reasoner and the Teacher with GRPO under the staged training scheme described in Section 3.4. Unless otherwise specified, all compared Reasoner methods share the same GRPO training pipeline and hyperparameters, differing in their reward designs. Experiments are conducted on two compute nodes, each equipped with 8 NVIDIA L40 GPUs (48GB): one for model optimization and the other for reward computation.

In the default T2R setup, the reward-service node hosts one frozen Teacher instance for reference CoT generation and multiple `Qwen3-4B-Instruct-2507` judge instances for pairwise CoT comparison and open-ended answer verification. Reward requests are processed through a unified Ray-based service with batched scheduling and inference. We apply KL regularization against the reference policy during GRPO training. The main hyperparameters are summarized in Table 7.

At each stage, the frozen Teacher $\tilde{\mathcal{T}}_i$ generates $K = 10$ reference CoTs per training sample using nucleus sampling with temperature 0.7 and top-$p = 0.95$. The maximum generation length is 4096 tokens, and the image preprocessing pixel range is set to $[262144, 524288]$. These settings are shared across all T2R stages.

Table 8 reports the Reasoner-side training cost under the same two-node setup, covering the shared R_0 stage and subsequent Reasoner training stages. Compared with RLVR Only, T2R requires $3.25\times$ and $3.03\times$ the wall-clock time at the 2B and 4B scales, respectively, mainly due to Teacher reference generation and pairwise CoT comparisons. Frozen Teacher incurs comparable overhead, while Judge Reward remains close to RLVR Only. The Teacher and judge are used only during training; deployment requires only the Reasoner.

B.2 Reasoner Prompt

The Reasoner prompt is instantiated from three placeholders: `{question}`, `{question_type}`, and `{question_type_desc}`. Here, `{question}` denotes the exam question, while `{question_type}` and `{question_type_desc}` specify the question type and the corresponding answer-format instruction.

For open-ended questions, `{question_type}` is set to `Open Ended Question`, and the answer-format instruction requires a short free-text answer. For single-choice questions, `{question_type}`

Table 7: Key hyperparameters used in staged GRPO training. Unless otherwise specified, these settings are shared across the Teacher, the Reasoner, and all comparison methods; differences arise only from the reward/objective design.

Category	Setting	Value
Algorithm	Objective	GRPO
	KL regularization	enabled
	KL coefficient	1×10^{-2}
Rollout	Rollouts per prompt	16
	Temperature	1.0
	Top-p	1.0
Data	Train rollout batch size	128
	Max prompt length	5120
	Max response length	1024
Optimization	Learning rate	1×10^{-6}
	Weight decay	1×10^{-2}
	Optimizer	AdamW (bf16)
	Max gradient norm	1.0
	Gradient checkpointing	enabled
	Vision tower freezing	disabled
	Padding-free training	enabled
	Dynamic batching	enabled
Training	Epochs per stage/split	1
System	RL training node	1 node, 8 × NVIDIA L40 (48GB)
	Reward-service node	1 node, 8 × NVIDIA L40 (48GB)
	Teacher reference instance	1 GPU instance
	Judge instances	7 GPU instances of `Qwen3-4B-Instruct-2507`
	Reward serving	unified Ray-based scheduling and request dispatch

Table 8: Reasoner-side training cost of the compared methods.

Method	2B		4B	
	Time (h)	GPU-h	Time (h)	GPU-h
RLVR Only	24	384	34	544
Judge Reward	25	400	36	576
Frozen Teacher	72	1,152	89	1,424
T2R	78	1,248	103	1,648

is set to `Single Choice Question`, and the answer-format instruction requires the model to output only the letter of the correct option, without brackets, quotation marks, or additional text.

These fields are inserted into a unified prompt template that requires the model to return a single XML response containing both a `<reason>` field and an `<answer>` field. The full prompt template is shown in Figure 8.

B.3 Teacher Prompt

The Teacher prompt is instantiated from five placeholders: `{question}`, `{question_type}`, `{question_type_desc}`, `{answer}`, and `{report}`. Here, `{question}` denotes the exam question, `{question_type}` and `{question_type_desc}` specify the question type and the corresponding answer-format instruction, `{answer}` is the official correct answer, and `{report}` is the radiology report for the same case.

Unlike the Reasoner, the Teacher is prompted to reconstruct a medically sound reasoning process while using the radiology report and the official answer only as alignment signals. The generated reasoning must be written as if the model had access only to the chest X-ray image and the exam question. Accordingly, the Teacher is explicitly instructed not to mention the radiology report, not

Prompt template for Reasoner

You will be given one or more chest X-ray images and an exam question. First, carefully inspect the X-ray image(s), then combine your observations with the question to reason step by step and produce a final answer.

Important requirements:
- In your reasoning, it is recommended (but not mandatory) to follow a natural sequence such as: "inspect the image(s) → describe key findings → analyze these findings in the context of the question → reach a conclusion". The exact wording can be fully free-form and does not need to follow any fixed template.
- The reasoning should clearly explain what you see on the image(s), what these findings imply, and how they lead step by step to your final answer. Do not invent imaging findings or test results that are not actually supported by the X-ray image(s).

Additional notes:
- Question type: {question_type}.
- The expected answer format for this type is described as follows (this is only to constrain the format of <answer>; you do not need to mention it in your reasoning):
{question_type_desc}

Question:
{question}

Based on what you see on the chest X-ray image(s) and the question, produce your response as a single XML structure in the following format:

```xml
<response>
  <reason>(fill in the complete reasoning here, in natural English as a continuous description)</reason>
  <answer>(fill in the final answer here, and make sure its format matches the requirement)</answer>
</response>
```

Do not output anything outside this XML code block.

Your Output:

Figure 8: Prompt template used for Reasoner training. The prompt includes the exam question, the question type, and the corresponding answer-format instruction, and requires the model to return a single XML response containing both `<reason>` and `<answer>`.

to quote the official answer, and not to present the reasoning as a reverse explanation of a known conclusion.

As in the Reasoner prompt, the question type and the corresponding answer-format instruction are included to align the generated reasoning with the downstream task setting. For open-ended questions, the prompt specifies a short free-text answer format; for single-choice questions, it specifies that the correct option should be represented by its letter only. These instructions are used only for task alignment and should not be explicitly referenced in the generated reasoning.

The Teacher is required to return a single XML response containing one `<response>` field, which stores the full reasoning trace in continuous natural English. The full prompt template is shown in Figure 9.

C Detailed Derivations for the Reasoner Objective

C.1 Non-Degenerate Case

In the non-degenerate case, the base reward already induces a non-trivial group-wise preference through

$$a_i^{\mathrm{raw}} = \tilde{r}_i^{\mathrm{rea}} - \frac{1}{G}\sum_{j=1}^{G} \tilde{r}_j^{\mathrm{rea}}, \tag{24}$$

which defines the subsets

$$\mathcal{P}^+ = \{i \in \mathcal{I} : a_i^{\mathrm{raw}} > 0\}, \qquad \mathcal{P}^- = \{i \in \mathcal{I} : a_i^{\mathrm{raw}} < 0\}. \tag{25}$$

In this case, the competition bonus is introduced only to refine the relative ordering within each subset; it should not alter the positive/negative partition induced by a_i^{raw}. We therefore verify that the proposed scaling preserves the sign of the base-reward advantage.

For $i \in \mathcal{P}^+$ and $i \in \mathcal{P}^-$, we define the subset-wise standardized patterns

$$p_i^+ = \frac{s_i^{\mathrm{rea}} - \mu_+^{\mathrm{rea}}}{\sigma_+^{\mathrm{rea}}}, \qquad p_i^- = \frac{s_i^{\mathrm{rea}} - \mu_-^{\mathrm{rea}}}{\sigma_-^{\mathrm{rea}}}. \tag{26}$$

We assume that the corresponding within-subset standard deviation is nonzero; when it vanishes, the competition bonus on that subset is set to zero.

Prompt template for Teacher

You will be given a chest X-ray exam question, the corresponding CXR (chest X-ray) image, a radiology report for the same case, and the official correct answer. Your task is to reconstruct a reasonable reasoning process for solving this question.

Important requirements:

- In the written reasoning, you must pretend that you only see the question and the chest X-ray itself. You do *not* see the radiology report, and you do *not* know the official answer in advance.
- The radiology report and the official answer are only for you to use "in the background": they help you keep your reasoning medically sound and ensure that your final conclusion is consistent with the official answer.
- Do **not** mention phrases like "the report says…", "according to the report…", "because the correct answer is…" in your reasoning, and do not make it look like you are reverse-explaining a known answer.
- In the reasoning, it is recommended (but not strictly required) to follow a natural sequence such as:
  "inspect the image → describe key findings → combine with the question to analyze → reach a conclusion".
  The exact wording and structure can be fully free-form; there is no fixed template.

Additional notes:

* Question type: {question_type}
* The expected answer format for this type is (for your internal alignment only; do not mention it in the reasoning):
  {question_type_desc}

Question:
{question}

Official correct answer (only for aligning your final conclusion; do not quote it in the reasoning text):
{answer}

Radiology report for this chest X-ray (only to help you understand the true case; do not explicitly quote or paraphrase it in the reasoning text):
{report}

Please write a complete reasoning process as if you were answering the question while reading the film, and output exactly one XML block in the following format:

```xml
<response>(write the full reasoning here in continuous natural English, without mentioning the report or the "official answer")</response>
```

Your output:

Figure 9: Prompt template used for Teacher training. The Teacher receives the exam question, question type, official correct answer, and radiology report, and is required to reconstruct a medically sound reasoning process while writing as if only the image and the question were observed. The output must be a single XML response containing a `<response>` field.

The competition bonus is

$$\Delta_i^{\mathrm{rea}} = \begin{cases} \eta_+^{\mathrm{rea}}\, p_i^+, & i \in \mathcal{P}^+, \\ \eta_-^{\mathrm{rea}}\, p_i^-, & i \in \mathcal{P}^-, \\ 0, & i \notin \mathcal{P}^+ \cup \mathcal{P}^-, \end{cases} \tag{27}$$

and the final advantage is

$$a_i^{\mathrm{final}} = a_i^{\mathrm{raw}} + \Delta_i^{\mathrm{rea}}. \tag{28}$$

To preserve the sign induced by the base reward, the scaling coefficients are defined as

$$\eta_+^{\mathrm{rea}} = \alpha \min_{i \in \mathcal{P}^+,\, p_i^+ < 0} \frac{a_i^{\mathrm{raw}}}{-p_i^+}, \qquad \eta_-^{\mathrm{rea}} = \alpha \min_{i \in \mathcal{P}^-,\, p_i^- > 0} \frac{-a_i^{\mathrm{raw}}}{p_i^-}, \tag{29}$$

where $\alpha \in (0, 1)$. If the constraining set $\{i \in \mathcal{P}^+ : p_i^+ < 0\}$ is empty, we set $\eta_+^{\mathrm{rea}} = 0$; similarly, if $\{i \in \mathcal{P}^- : p_i^- > 0\}$ is empty, we set $\eta_-^{\mathrm{rea}} = 0$.

We now verify that this construction preserves the sign of a_i^{raw}.

Positive subset. Let $i \in \mathcal{P}^+$. Then $a_i^{\mathrm{raw}} > 0$. If $p_i^+ \geq 0$, we have

$$a_i^{\mathrm{final}} = a_i^{\mathrm{raw}} + \eta_+^{\mathrm{rea}} p_i^+ \geq a_i^{\mathrm{raw}} > 0, \tag{30}$$

so the sign is preserved automatically.

It remains to consider the constraining case $p_i^+ < 0$. Since

$$\min_{j \in \mathcal{P}^+,\, p_j^+ < 0} \frac{a_j^{\mathrm{raw}}}{-p_j^+} \leq \frac{a_i^{\mathrm{raw}}}{-p_i^+}, \tag{31}$$

it follows that

$$\eta_+^{\mathrm{rea}} = \alpha \min_{j \in \mathcal{P}^+,\, p_j^+ < 0} \frac{a_j^{\mathrm{raw}}}{-p_j^+} \leq \alpha \frac{a_i^{\mathrm{raw}}}{-p_i^+}. \tag{32}$$

Because $p_i^+ < 0$, multiplying both sides by p_i^+ reverses the inequality:

$$\eta_+^{\text{rea}} p_i^+ \geq -\alpha a_i^{\text{raw}}. \tag{33}$$

Hence,

$$a_i^{\text{final}} = a_i^{\text{raw}} + \eta_+^{\text{rea}} p_i^+ \geq (1-\alpha) a_i^{\text{raw}} > 0, \tag{34}$$

where the strict inequality follows from $\alpha \in (0,1)$ and $a_i^{\text{raw}} > 0$.

Negative subset. Let $i \in \mathcal{P}^-$. Then $a_i^{\text{raw}} < 0$. If $p_i^- \leq 0$, we have

$$a_i^{\text{final}} = a_i^{\text{raw}} + \eta_-^{\text{rea}} p_i^- \leq a_i^{\text{raw}} < 0, \tag{35}$$

so the sign is again preserved automatically.

Now consider the constraining case $p_i^- > 0$. Since

$$\min_{j \in \mathcal{P}^-,\, p_j^- > 0} \frac{-a_j^{\text{raw}}}{p_j^-} \leq \frac{-a_i^{\text{raw}}}{p_i^-}, \tag{36}$$

we obtain

$$\eta_-^{\text{rea}} = \alpha \min_{j \in \mathcal{P}^-,\, p_j^- > 0} \frac{-a_j^{\text{raw}}}{p_j^-} \leq \alpha \frac{-a_i^{\text{raw}}}{p_i^-}. \tag{37}$$

Since $p_i^- > 0$, multiplying by p_i^- preserves the inequality:

$$\eta_-^{\text{rea}} p_i^- \leq -\alpha a_i^{\text{raw}}. \tag{38}$$

Therefore,

$$a_i^{\text{final}} = a_i^{\text{raw}} + \eta_-^{\text{rea}} p_i^- \leq (1-\alpha) a_i^{\text{raw}} < 0, \tag{39}$$

again because $\alpha \in (0,1)$ and $a_i^{\text{raw}} < 0$.

Combining the two subset analyses, we conclude that the competition bonus refines the within-subset ordering without changing the sign induced by the base reward. In particular, $a_i^{\text{final}} > 0$ for all $i \in \mathcal{P}^+$, $a_i^{\text{final}} < 0$ for all $i \in \mathcal{P}^-$, and $a_i^{\text{final}} = 0$ for all $i \notin \mathcal{P}^+ \cup \mathcal{P}^-$. Thus, the proposed bonus is sign-preserving and cannot alter the positive/negative assignment induced by the base reward.

C.2 Degenerate Case

In the degenerate case, all samples in the group receive the same base reward, so

$$a_i^{\text{raw}} = 0, \qquad \forall i \in \mathcal{I}. \tag{40}$$

Hence, the base reward induces no non-trivial group-wise preference, and the final advantage must be reconstructed directly from the competition scores.

For a candidate threshold τ, we define

$$\mathcal{P}^-(\tau) = \{i \in \mathcal{I} : s_i^{\text{rea}} \leq \tau\}, \qquad \mathcal{P}^+(\tau) = \{i \in \mathcal{I} : s_i^{\text{rea}} > \tau\}. \tag{41}$$

We consider candidate thresholds drawn from the ordered values of $\{s_i^{\text{rea}}\}_{i=1}^G$. Among all thresholds that induce non-empty subsets on both sides, we choose the largest feasible τ satisfying

$$\sum_{i \in \mathcal{P}^-(\tau)} s_i^{\text{rea}} < \gamma \sum_{i \in \mathcal{P}^+(\tau)} s_i^{\text{rea}}, \tag{42}$$

where $\gamma > 0$ is a hyperparameter. If no feasible threshold exists, we set

$$a_i^{\text{final}} = 0, \qquad \forall i \in \mathcal{I}. \tag{43}$$

We therefore consider only the feasible case, which yields induced subsets $\mathcal{P}^+$ and $\mathcal{P}^-$.

We first globally center the competition scores:

$$p_i = s_i^{\text{rea}} - \mu^{\text{rea}}, \qquad \mu^{\text{rea}} = \frac{1}{G} \sum_{j=1}^{G} s_j^{\text{rea}}. \tag{44}$$

By construction,

$$\sum_{i=1}^{G} p_i = 0. \tag{45}$$

To obtain a reconstructed signal that is both zero-mean and sign-consistent with the induced partition, we shift only the positive subset:

$$\hat{p}_i = \begin{cases} p_i + \lambda, & i \in \mathcal{P}^+, \\ p_i, & i \in \mathcal{P}^-. \end{cases} \tag{46}$$

Let

$$m = |\mathcal{P}^+|, \qquad n = |\mathcal{P}^-|, \qquad m + n = G. \tag{47}$$

The group mean of $\{\hat{p}_i\}_{i=1}^{G}$ is then

$$\frac{1}{G}\sum_{i=1}^{G} \hat{p}_i = \frac{1}{G}\left(\sum_{i=1}^{G} p_i + m\lambda\right) = \frac{m\lambda}{G}. \tag{48}$$

The final advantage is defined as

$$a_i^{\text{final}} = \hat{p}_i - \frac{1}{G}\sum_{j=1}^{G} \hat{p}_j. \tag{49}$$

To ensure consistency with the induced partition, we require

$$a_i^{\text{final}} \geq 0, \qquad \forall i \in \mathcal{P}^+, \tag{50}$$

and

$$a_i^{\text{final}} \leq 0, \qquad \forall i \in \mathcal{P}^-. \tag{51}$$

Positive subset constraint. For $i \in \mathcal{P}^+$,

$$a_i^{\text{final}} = \hat{p}_i - \frac{1}{G}\sum_{j=1}^{G} \hat{p}_j = p_i + \lambda - \frac{m\lambda}{G} = p_i + \frac{n}{G}\lambda. \tag{52}$$

Hence, the non-negativity constraint reduces to

$$p_i + \frac{n}{G}\lambda \geq 0 \quad \Longleftrightarrow \quad \lambda \geq -\frac{G}{n}p_i. \tag{53}$$

To satisfy this for all $i \in \mathcal{P}^+$, it is sufficient to require

$$\lambda \geq -\frac{G}{n}\min_{i\in\mathcal{P}^+} p_i. \tag{54}$$

Negative subset constraint. For $i \in \mathcal{P}^-$,

$$a_i^{\text{final}} = \hat{p}_i - \frac{1}{G}\sum_{j=1}^{G} \hat{p}_j = p_i - \frac{m\lambda}{G}. \tag{55}$$

Hence, the non-positivity constraint reduces to

$$p_i - \frac{m\lambda}{G} \leq 0 \quad \Longleftrightarrow \quad \lambda \geq \frac{G}{m}p_i. \tag{56}$$

To satisfy this for all $i \in \mathcal{P}^-$, it is sufficient to require

$$\lambda \geq \frac{G}{m}\max_{i\in\mathcal{P}^-} p_i. \tag{57}$$

Combining the two subset constraints, the minimum feasible shift is

$$\lambda_{\min} = \max\left(-\frac{G}{|\mathcal{P}^-|}\min_{i\in\mathcal{P}^+} p_i,\ \frac{G}{|\mathcal{P}^+|}\max_{i\in\mathcal{P}^-} p_i\right), \tag{58}$$

and we set

$$\lambda = \max(\lambda_{\min}, 0). \tag{59}$$

With this choice of λ, the reconstructed final advantage is zero-mean at the group level and sign-consistent with the partition induced by the competition scores. It therefore provides a valid replacement training signal when the original group-level advantage collapses to zero.

Table 9: Additional results with `InternVL3-2B` as the base VLM.

Method	Med-CXR	VQA-RAD	SLAKE	MIMIC-CXR	Covid19	CXR-PA	Avg.
Base Model	23.55	35.60	47.64	15.69	16.98	22.36	26.97
RLVR Only	23.32	35.96	48.79	16.94	18.30	22.07	27.56
T2R-R1	24.43	39.12	52.11	19.26	29.72	28.82	32.24
T2R-R2	28.61	40.90	**53.54**	21.18	33.16	32.67	35.01
T2R-R3	**30.35**	**43.49**	50.52	**22.19**	**38.83**	**36.20**	**36.93**

D Additional Experiments

D.1 Additional Backbone Results

Our main experiments use Qwen3-VL as the base VLM. To examine whether the improvements of T2R extend to a different model family, we further conduct three-stage T2R training with `InternVL3-2B` [50]. We follow the same training and evaluation protocol described in Section 4.1, replacing only the base VLM.

Table 9 shows that the average accuracy of T2R increases across the three training stages. The final-stage T2R-R3 outperforms both the Base Model and RLVR Only on all six benchmarks, with average gains of 9.96 and 9.37 percentage points, respectively. These results demonstrate that the improvements of T2R extend from Qwen3-VL to InternVL3, supporting its applicability across different VLM backbones for CXR VQA.

D.2 Evaluation Stability across Judge Runs

As described in Section 4.1, we evaluate each prediction five times independently using Qwen3-235B-A22B and determine its final correctness by per-sample majority vote. To examine evaluation stability, we also compute accuracy separately for each judge run and report the mean and standard deviation in Table 10.

Table 10 shows that the five-run mean accuracies broadly follow the trends observed in the majority-vote results. At 2B, T2R-R3 achieves the highest mean accuracy on five of the six benchmarks, with SLAKE as the exception. At 4B, T2R-R2 shows notable improvements over Frozen Teacher on VQA-RAD and Covid19, while their results on SLAKE are close (64.92 ± 0.64 vs. 64.42 ± 0.84). Similarly, T2R-R3 performs comparably to Frozen Teacher on VQA-RAD (43.37 ± 0.46 vs. 43.16 ± 0.72), but underperforms it on SLAKE.

Table 10: Accuracy (%, mean $\pm$ standard deviation) across five independent final-answer judge runs. Best mean accuracies within each model scale and dataset are shown in bold. Rows corresponding to T2R are shaded in light gray.

Scale	Method	Datasets					
		Med-CXR	VQA-RAD	SLAKE	MIMIC-CXR	Covid19	CXR-PA
2B	Base Model	17.85 ± 0.09	32.72 ± 0.91	52.72 ± 1.07	15.06 ± 0.23	24.13 ± 0.60	32.75 ± 0.57
	RLVR Only	18.53 ± 0.07	31.18 ± 0.86	53.73 ± 1.02	13.76 ± 0.22	23.05 ± 0.56	21.80 ± 0.54
	Judge Reward	28.97 ± 0.06	36.47 ± 0.82	$\mathbf{55.76 \pm 0.96}$	18.96 ± 0.20	17.21 ± 0.53	28.05 ± 0.51
	Frozen Teacher	22.70 ± 0.04	40.70 ± 0.72	52.28 ± 0.84	21.87 ± 0.18	32.50 ± 0.47	32.77 ± 0.45
	T2R-R1	23.16 ± 0.06	41.34 ± 0.60	52.30 ± 0.71	21.26 ± 0.15	39.68 ± 0.39	35.02 ± 0.38
	T2R-R2	27.96 ± 0.05	40.64 ± 0.54	53.84 ± 0.64	22.73 ± 0.14	31.00 ± 0.36	37.25 ± 0.34
	T2R-R3	$\mathbf{31.99 \pm 0.04}$	$\mathbf{43.42 \pm 0.46}$	51.77 ± 0.54	$\mathbf{24.24 \pm 0.12}$	$\mathbf{40.52 \pm 0.30}$	$\mathbf{37.75 \pm 0.29}$
4B	Base Model	23.32 ± 0.11	40.92 ± 0.91	57.66 ± 1.07	22.11 ± 0.23	27.61 ± 0.59	41.72 ± 0.57
	RLVR Only	22.77 ± 0.07	41.58 ± 0.87	56.72 ± 1.02	24.04 ± 0.22	29.90 ± 0.56	36.53 ± 0.54
	Judge Reward	23.60 ± 0.06	41.52 ± 0.82	57.25 ± 0.96	24.28 ± 0.20	30.08 ± 0.53	37.00 ± 0.51
	Frozen Teacher	$\mathbf{25.28 \pm 0.06}$	43.16 ± 0.72	64.42 ± 0.84	26.68 ± 0.18	28.33 ± 0.47	38.98 ± 0.45
	T2R-R1	21.96 ± 0.05	37.77 ± 0.60	60.99 ± 0.71	20.55 ± 0.15	40.77 ± 0.39	42.44 ± 0.38
	T2R-R2	23.41 ± 0.04	$\mathbf{46.86 \pm 0.54}$	$\mathbf{64.92 \pm 0.64}$	26.11 ± 0.14	$\mathbf{49.37 \pm 0.36}$	$\mathbf{45.42 \pm 0.34}$
	T2R-R3	23.61 ± 0.04	43.37 ± 0.46	60.58 ± 0.54	$\mathbf{28.35 \pm 0.12}$	41.48 ± 0.30	43.96 ± 0.29

Table 11: Preference rates (%) of updated Teacher CoTs over those from $\mathcal{T}_0$. Each row evaluates 1,000 CoT pairs matched by image–question instance; 50% indicates no systematic preference.

Teacher Comparison	Factuality	Reasoning	Answer Support
$\mathcal{T}_1$ vs. $\mathcal{T}_0$	72.6	83.3	82.8
$\mathcal{T}_2$ vs. $\mathcal{T}_0$	78.6	87.5	89.2
$\mathcal{T}_3$ vs. $\mathcal{T}_0$	**82.1**	**92.0**	**94.7**

Table 12: Comparison accuracy and agreement (%) under controlled factuality conditions. Each condition contains 1,000 CoT pairs.

Judge	Both Correct Agreement	One Correct Selection Accuracy	Both Incorrect Agreement
Qwen3-4B-Instruct-2507	80.9	71.4	81.6
Qwen3-8B	85.8	76.6	87.5
Qwen3-32B	**91.1**	**80.4**	**90.9**

D.3 Independent Evaluation of Teacher Improvement

The Teacher in T2R generates reference CoTs during training rather than serving as the final VQA predictor. Since it receives the radiology report and reference answer in addition to the image and question, standalone Teacher VQA accuracy would be confounded by these privileged inputs. We therefore directly evaluate the quality of its generated reference CoTs.

We use a fixed set of 1,000 image–question instances and generate one CoT from each Teacher $\mathcal{T}_0$, $\mathcal{T}_1$, $\mathcal{T}_2$, and $\mathcal{T}_3$ under the same decoding configuration. For each updated Teacher $\mathcal{T}_i$, we compare its CoTs with those from $\mathcal{T}_0$ on the same instances along three dimensions: *medical factuality*, assessing the correctness of medical facts and imaging findings; *reasoning correctness*, assessing the correctness and internal consistency of the reasoning; and *answer support*, assessing whether the CoT provides sufficient and relevant evidence for the reference answer.

We use Baichuan-M2-32B [51] to evaluate medical factuality and Qwen3-235B-A22B to evaluate reasoning correctness and answer support. Neither model participates in T2R training. The A/B order is randomized for each comparison, and the evaluator is required to select one winner. Table 11 reports the percentage of comparisons in which the updated Teacher's CoT is preferred over that from $\mathcal{T}_0$ across the three evaluation dimensions.

The preference rates increase monotonically across all three dimensions as self-competition progresses. For $\mathcal{T}_3$, the preference rates reach 82.1% for medical factuality, 92.0% for reasoning correctness, and 94.7% for answer support. These results provide independent evidence that self-competition progressively improves the quality of Teacher-generated reference CoTs.

D.4 Reliability of the Comparison Judge

The pairwise comparison judge provides the preference signal used in T2R training. To assess its reliability, we evaluate judges of different model scales on a controlled set of CoT pairs generated during T2R training.

We use Baichuan-M2-32B [51], which is not involved in T2R training, to determine whether each CoT contains medical factual errors based on the radiology report, question, and reference answer. We divide the pairs into three conditions: both CoTs are factually correct, only one CoT is factually correct, and both CoTs are factually incorrect. We randomly sample 1,000 pairs from each condition.

For pairs with only one factually correct CoT, we report how often the judge selects the correct one. When both CoTs share the same factuality status, factual correctness alone cannot distinguish their relative quality. We therefore evaluate agreement with the overall reasoning-quality preferences of Qwen3-235B-A22B, which is also independent of T2R training. The A/B order is randomized for all comparisons. Table 12 reports the selection accuracy or agreement under each condition.

The results show that judge performance improves consistently with model scale across all three conditions. When only one CoT is factually correct, selection accuracy increases from 71.4% for Qwen3-4B-Instruct-2507 to 76.6% and 80.4% for Qwen3-8B and Qwen3-32B, respectively.

Table 13: Agreement (%) of different comparison judges with the reference winner under different relative CoT lengths. Each group contains 1,000 pairs.

Judge	Better Shorter	Similar Length	Better Longer
Qwen3-4B-Instruct-2507	67.2	78.4	88.3
Qwen3-8B	76.0	84.7	94.5
Qwen3-32B	**87.1**	**92.2**	**97.1**

Agreement with the external reference judge also improves when both CoTs share the same factuality status. These findings indicate that the original 4B judge provides a useful but imperfect comparison signal, while larger judges offer more reliable pairwise comparisons.

D.5 Length Sensitivity of the Comparison Judge

The comparison prompt instructs the judge to evaluate CoTs based on their content rather than their length. We further investigate whether the resulting judgments are nevertheless sensitive to relative CoT length.

We construct a separate set of CoT pairs and determine a reference winner for each pair using Baichuan-M2-32B and Qwen3-235B-A22B, following the evaluation protocol in Appendix D.4. For each pair, we define the relative length difference as

$$\Delta L = \frac{L_{\text{winner}} - L_{\text{other}}}{L_{\text{other}}}, \tag{60}$$

where L denotes the CoT token count and winner denotes the CoT preferred by the reference evaluation. We divide the pairs into three groups: the reference winner is shorter ($\Delta L < -10\%$), the two CoTs have similar length ($-10\% \leq \Delta L \leq 10\%$), or the reference winner is longer ($\Delta L > 10\%$). We randomly sample 1,000 pairs from each group and randomize the A/B order in all comparisons.

We evaluate `Qwen3-4B-Instruct-2507`, `Qwen3-8B`, and `Qwen3-32B` as comparison judges on the same pairs. Table 13 reports the percentage of comparisons in which each judge selects the reference winner under the three length conditions.

The results reveal a clear length-associated asymmetry in the original 4B judge. Its agreement with the reference evaluation is 67.2% when the reference winner is shorter, compared with 88.3% when it is longer, a gap of 21.1 percentage points. This gap decreases to 18.5 points for the 8B judge and 10.0 points for the 32B judge. Agreement with the reference evaluation also increases across all three length conditions as judge scale increases. These results indicate that larger judges provide more reliable comparisons with smaller length-associated differences.

E Qualitative Case Studies

We provide qualitative examples in Figure 10–12 to compare the reasoning traces generated by **RLVR Only**, **Judge Reward**, **Frozen Teacher**, and **T2R-R3**. Across these examples, different supervision signals are associated with visibly different CoT structures. **RLVR Only** tends to produce answer-oriented rationales with limited intermediate analysis, which is consistent with its lack of explicit CoT-level supervision. **Judge Reward** improves local fluency and coherence, but the resulting reasoning often remains relatively brief and stops after identifying a few relevant findings. **Frozen Teacher** further encourages more organized rationales, including explicit summaries of key findings. In contrast, **T2R-R3** more consistently follows a complete reasoning pattern: it first analyzes the chest X-ray findings, then summarizes the relevant evidence, evaluates the candidate options when applicable, and finally derives the answer from the accumulated observations.

This qualitative pattern is consistent with the training signal provided by T2R. Directly judging a single CoT with an LLM can encourage the model to satisfy the predefined acceptability criteria in the judge prompt (Figure 7), but such supervision may become less informative as the model adapts to these criteria. By contrast, T2R relies on pairwise comparison against reference CoTs produced by a Teacher that is itself improved through self-competition. As the reference CoTs become stronger, the Reasoner continues to receive relative preference signals that favor more grounded, structured,

and option-aware reasoning. These examples therefore suggest that comparison-based supervision can improve not only final answer prediction, but also the organization and reliability of the generated CoT.

F Limitations and Future Work

The effectiveness of T2R depends on the reliability of its pairwise comparison judge. Our analyses show that the original 4B judge has limited accuracy in distinguishing factually correct from incorrect CoTs, and that its agreement with independent reference evaluations varies with relative CoT length (Appendices D.4 and D.5). Such limitations may introduce noisy or biased supervision during training. Although stronger judges improve comparison reliability and downstream performance, developing more robust comparison signals remains an important direction.

T2R also incurs additional training-time computation for Teacher reference-CoT generation and pairwise comparisons. This increases training costs compared with standard RLVR, even with our separate reward-service node and Ray-based request dispatch. Improving the efficiency of comparison-based training therefore remains important. The Teacher and judge are used only during training, leaving the deployed Reasoner unchanged.

Another limitation is that the current comparison judge evaluates multiple aspects of CoT quality through a single overall preference. While this provides a simple supervision signal, it does not explicitly distinguish the contributions of individual criteria, such as medical factuality, evidence grounding, and answer support. Future work could introduce finer-grained, criterion-specific comparisons to provide more interpretable and controllable supervision.

Finally, our empirical evaluation focuses on CXR VQA using Qwen3-VL and InternVL3 backbones. Exploring whether T2R extends to other reasoning tasks and structured outputs, such as planning traces and code, remains a direction for future work.

G Broader Impacts

This work studies reinforcement-learning-based reasoning optimization for chest X-ray visual question answering. Its potential positive impact is to improve the reliability and interpretability of medical multimodal models by encouraging more evidence-grounded and coherent reasoning traces. Such techniques may be useful for developing decision-support systems, medical education tools, or research assistants that provide more transparent intermediate reasoning. At the same time, this work also has potential risks. Models trained with T2R may still produce incorrect answers or medically implausible rationales, and LLM-as-a-Judge supervision may introduce biases or failure modes from the judge model. In clinical settings, incorrect or overconfident reasoning could negatively affect decision-making if used without expert oversight. Therefore, the models and methods studied in this paper should not be used as standalone clinical decision systems. Any practical deployment would require careful validation on diverse populations, monitoring for bias and failure cases, privacy-preserving data handling, and use under the supervision of qualified medical professionals.

CoT Examples





What is the most likely diagnosis based on the chest X-ray images?

A. Rib fracture
B. Bone lesion
C. None of the above
D. Pneumothorax

RLVR Only

Upon inspection of the chest X-ray images, the radiographic appearance shows a normal chest with no visible signs of a fracture in the ribs. There are no signs of a bone lesion, which would be indicated by an abnormality in the bone structure. The lung fields appear clear with no evidence of air in the pleural space, which would be a sign of pneumothorax. The image does not show any evidence of a rib fracture, as there are no breaks or deformities in the ribs. Therefore, the most likely diagnosis based on the chest X-ray images is that there is no fracture or lesion.

Judge Reward

Inspecting the chest X-ray images, there are no visible signs of a rib fracture, such as a broken or displaced rib, which would typically show a fracture line or a deformity in the bone. There is also no evidence of a pneumothorax, which would be indicated by a visible air collection in the pleural space, which is not apparent in the image. Furthermore, there is no visible bone lesion, such as a mass or a focal abnormality in the bone. The image appears to show a normal chest cavity with no signs of air in the pleural space or a fracture. The structures of the ribs and the lungs appear intact and are not showing any signs of abnormality.

Frozen Teacher

Inspecting the chest X-ray image(s), there are no visible signs of rib fractures, such as broken or deformed ribs, which would be indicated by fractures or angulation. There is no evidence of a pneumothorax, as the lung fields appear clear and there are no signs of air in the pleural space (such as a collapsed lung or a visible shift of the mediastinum). The image does not show any abnormal bone density or lesion that would suggest a bone lesion, such as a tumor or osteolytic change. The radiographic appearance is consistent with normal anatomy and no pathological findings are present. Therefore, the most likely diagnosis based on the X-ray findings is none of the listed conditions. The absence of any structural abnormalities in the ribs, lungs, or pleural space rules out rib fracture, bone lesion, and pneumothorax. This leads to the conclusion that the most appropriate answer is "None of the above" because none of the listed conditions are supported by the image findings.

Key findings from the X-ray:
- No visible rib fractures: The ribs appear intact with normal spacing and no signs of breakage or displacement.
- No pneumothorax: The lung fields are clear and the costophrenic angles are sharp, indicating no air in the pleural space.
- No bone lesion: There is no evidence of a mass, osteolytic changes, or abnormal bone density.
- Normal anatomical structures: The heart shadow and diaphragm are in normal positions with no abnormal shift.

Given that none of the listed conditions (A: Rib fracture, B: Bone lesion, D: Pneumothorax) are supported by the radiographic findings, the correct answer is C: None of the above.

T2R-R3

Inspecting the chest X-ray image(s), the most critical observation is the complete absence of any abnormal radiopaque (white) or radiolucent (dark) density within the bony thoracic structures—specifically, the ribs, clavicles, and vertebrae. There are no fractures, no lucent (dark) areas indicating bone destruction, and no discontinuity or separation of any rib margin. All ribs appear intact, with smooth, continuous cortical margins and normal radiopacity throughout their entire length. The costochondral junctions (where ribs meet the costal cartilage) are unremarkable—no displacement, separation, or periosteal reaction is visible. This is a defining feature of a "normal" chest X-ray: no structural disruption of the bony thorax.

Key findings:
- **No rib fractures**: No breakage, no lucent lines, no separated or angulated ribs. Every rib shows a complete, continuous cortical outline. This is inconsistent with rib fracture, which would manifest as either a lucent line (gap) along a rib segment or a radiolucent fracture line (e.g., a crack or separation).
- **No bone lesion**: There is no localized area of increased or decreased opacity (e.g., sclerotic or lytic lesion), no periosteal reaction, no cortical thinning, and no bone destruction or abnormal bone formation (e.g., osteolytic or osteoblastic changes). The bone density is uniform throughout the thoracic skeleton. This rules out bone lesion, which would show either:
  - A localized radiodense area (sclerosis) or
  - A radiolucent area (lysis) in the bone cortex.
- **No pneumothorax**: The costophrenic angles (where the diaphragm meets the costal margin) are sharp and well-defined. There is no visible visceral pleural line with absence of peripheral lung markings, which is the definitive radiographic sign of pneumothorax. The diaphragm is intact and shows normal contour, with no flattening, elevation, or deep sulcus sign. This confirms the absence of pneumothorax.

Now evaluating the answer choices:
- **A: Rib fracture** → Clearly contradicted. Rib fractures produce radiolucent lines along the bone cortex (visible as a break or gap), which are absent here. The ribs are uniformly radiopaque with no breaks or separation. Therefore, A is definitively incorrect.
- **B: Bone lesion** → Also definitively ruled out. A bone lesion would produce either:
  - A localized radiodense area (sclerosis) or
  - A radiolucent area (lysis) in the cortex.

  Neither is present. The entire thoracic bone architecture is normal in density and continuity. Thus, B is incorrect.
- **C: None of the above** → This is the only option that remains consistent with the imaging findings. Since both A and B are directly falsified by the X-ray, and no other diagnosis is supported by the image, this is the only logically valid conclusion. The image shows no structural abnormality in the bony thorax, which is precisely what "none of the above" describes.
- **D: Pneumothorax** → This is explicitly ruled out by the absence of the radiographic hallmark: **a visible visceral pleural line with absent peripheral lung markings**. The costophrenic angles are sharp and well-defined, and there is no visible pleural line, no absent peripheral lung markings, no pleural effusion, and no deep sulcus sign. Therefore, D is definitively incorrect.

In summary, all clinical and radiographic criteria point unambiguously to a completely normal thoracic skeletal anatomy. The absence of any radiolucent or radiodense changes in the ribs, clavicles, and vertebrae, combined with the absence of a visible pleural line or peripheral loss of lung markings, is incompatible with rib fracture, bone lesion, or pneumothorax. Thus, the only possible diagnosis consistent with the X-ray findings is "None of the above."

Figure 10: Qualitative comparison of CoTs generated by different models.

CoT Examples



What specific finding is observed in the lungs on this chest X-ray image?

A. Cardiomegaly
B. Pulmonary consolidation
C. Pleural effusion
D. Slight hyperaeration

RLVR Only

Upon inspection of the chest X-ray image, the lung fields appear clear and well-defined. There are no signs of abnormal shadows, masses, or fluid accumulation in the pleural space. The lung markings are visible and appear normal, without evidence of consolidation or opacification. The heart size appears within normal limits, and there is no obvious enlargement. The diaphragm is visible and appears to be in a normal position. The overall lung appearance is consistent with normal aerated lung tissue. There is no evidence of hyperinflation, which would typically be seen in conditions like COPD or asthma. The image does not show signs of pleural effusion, such as a blunting of the costophrenic angles or a fluid-filled space. The presence of a slight hyperaeration is not indicated by the image, as there are no areas of increased opacity or darkening in the lung fields. The image does not show any signs of cardiomegaly, which would be indicated by an enlarged heart shadow. Therefore, the most likely finding is a normal appearance of the lungs with no specific abnormalities.

Judge Reward

Inspecting the chest X-ray image, the lungs appear to have a relatively clear view with no obvious signs of consolidation, effusion, or significant mass. The lung fields are visible with no apparent dark areas that would suggest pulmonary consolidation. The diaphragm is in its normal position, and the heart appears to be within normal limits. There is a subtle appearance of increased lung transparency, particularly in the lower and upper lung zones, which could be indicative of slight hyperaeration. This is a common finding in some patients with mild respiratory conditions, such as those with asthma or other chronic conditions, but it is not a severe condition. The bronchovascular markings are visible and do not show signs of significant narrowing or thickening. There is no evidence of pleural effusion, which would typically present as a fluid level or blunting of the costophrenic angles. The heart is not enlarged, and there are no signs of cardiomegaly. Therefore, the most likely finding is slight hyperaeration.

Frozen Teacher

Inspecting the chest X-ray image, the lung fields appear to have increased radiolucency, particularly in the lower lobes, which is indicative of a relatively normal or slightly increased air content in the lungs. There is no evidence of consolidation (which would show as a dense, white area), pleural effusion (which would appear as a blunting of the costophrenic angle or fluid in the pleural space), or significant cardiomegaly (which would be visible as an enlarged heart silhouette). The most prominent observation is the slight increase in the air content of the lungs, which is consistent with a "slight hyperaeration" – a term used to describe a mild increase in lung volume or air content, typically seen in the absence of significant disease.

Key findings on the image include:
- The lung fields show a subtle increase in radiolucency, especially in the lower zones, which is characteristic of slight hyperaeration.
- There is no evidence of a dense opacity, which would suggest consolidation.
- The heart size is within normal limits, with no apparent enlargement (i.e., no cardiomegaly).
- The costophrenic angles are sharp and clear, indicating no pleural effusion.

These findings are consistent with a "slight hyperaeration," which is a term used to describe a mild increase in lung volume or air content, typically seen in the absence of significant disease. This is the only finding that is directly and clearly observed on the chest X-ray image and aligns with the question.

When analyzing the options:
- A: Cardiomegaly is not present; the heart appears normal in size.
- B: Pulmonary consolidation is not present; there is no dense, white area in the lung fields.
- C: Pleural effusion is not present; the costophrenic angles are sharp and clear.
- D: Slight hyperaeration matches the observed radiolucency in the lungs, particularly in the lower lobes, and is the most accurate description of the finding.

Therefore, the specific finding observed in the lungs is "slight hyperaeration."

T2R-R3

Inspecting the chest X-ray image, the most prominent and consistent finding is the **slight increase in lung transparency (hyperaeration)**, specifically visible as a **mild increase in radiolucency** in the lung fields—particularly noticeable in the lower lobes and costophrenic angles. This is best described as “slight hyperaeration,” which is characterized by:
- **Preserved costophrenic angle sharpness** (i.e., the angle between the diaphragm and the chest wall is not blunted), which is preserved in the lower zones.
- **Increased radiolucency** (lower density) in the lung fields, especially in the lower lung zones, without blunting of the costophrenic margins or the presence of pleural fluid levels.
- **No blunting of costophrenic or costodiaphragmatic angles**, which would be expected in pleural effusion, while consolidation would more typically appear as focal or lobar increased opacity within the lung field.

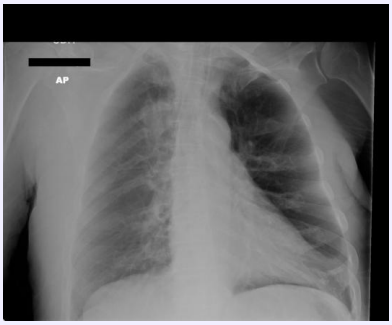
Now, evaluating the choices:
- **A: Cardiomegaly** – This would manifest as an enlarged cardiac silhouette, particularly an increased cardiothoracic ratio. The cardiac borders are **not enlarged** or displaced; the cardiothoracic ratio is **within normal limits**. No cardiac enlargement is seen. Thus, A is ruled out.
- **B: Pulmonary consolidation** – This would present as **focal or lobar increased opacity in the lung parenchyma**, often with **loss of normal radiolucency** and sometimes with **air bronchograms**. The image clearly shows **no focal air-space opacity**; instead, the lung fields remain relatively radiolucent and the costophrenic margins remain sharp and well-defined. This is a key distinguishing feature: **no focal opacity = not consolidation**. Therefore, B is incorrect.
- **C: Pleural effusion** – This would cause **blunting of the costophrenic or costodiaphragmatic angles** (typically at least one is blunted), often with **loss of sharpness** of the costophrenic margins or a meniscus-like basal opacity. The image shows **no blunting**—the costophrenic angles remain clearly defined and sharp, and there is no pleural fluid level or basal opacity. Therefore, C is ruled out.
- **D: Slight hyperaeration** – This exactly matches the observed radiographic pattern: **increased lung transparency (radiolucency)** in the lower lung zones, with **preserved costophrenic angles** and **no blunting**. This is consistent with slight hyperaeration: the lung parenchyma is aerated but slightly more radiolucent than normal, with no structural distortion. All other features (e.g., preserved costophrenic angles, no fluid levels, no blunting) are consistent with this diagnosis.

The entire radiographic pattern—**preserved costophrenic angles, increased radiolucency in lower lung zones, and no blunting**—is most consistent with **slight hyperaeration**. None of the other choices produce a radiographic pattern that matches the image as well. The other choices (cardiomegaly, consolidation, pleural effusion) are clinically and radiographically not favored by the findings shown.

Therefore, the correct and best-supported interpretation of the image is **D: Slight hyperaeration**.

Figure 11: Qualitative comparison of CoTs generated by different models.

CoT Examples



Given the presence of a right upper lobe mass and the clinical history, what is the most likely diagnosis?

A. Primary lung cancer
B. Pneumonia
C. Pulmonary embolism
D. Tuberculosis

RLVR Only

The chest X-ray image shows a distinct opacity in the right upper lobe, which is consistent with a mass lesion. The presence of a mass in the right upper lobe is a key finding that would suggest a primary lung tumor. The lesion appears as a well-defined, dense area with a clear boundary, which is typical of a primary lung cancer. The clinical history, although not fully detailed, is likely to include symptoms such as cough, chest pain, or weight loss, which are common in lung cancer. Other options such as pneumonia or tuberculosis typically present with different radiographic features (e.g., consolidation or infiltrate, or specific patterns on imaging), and pulmonary embolism is typically associated with signs of vascular obstruction, which are not visible here. Therefore, the most likely diagnosis is primary lung cancer.

Judge Reward

Inspecting the chest X-ray image, there is a visible mass in the right upper lobe. This finding is significant because it is a key diagnostic feature. The presence of a mass in the right upper lobe is not consistent with typical presentations of pneumonia or pulmonary embolism, which are usually characterized by different patterns such as consolidation or signs of vascular occlusion, respectively. Tuberculosis, while it can present with a mass, typically shows more granulomatous changes and often has a history of chronic infection and exposure to mycobacteria. However, the clinical history is important in differentiating between these conditions. The fact that the mass is located in the right upper lobe suggests a possible malignancy, as this region is a common site for primary lung cancers. The appearance of the mass is not consistent with the typical features of infection or embolism. Therefore, given the clinical context and the imaging findings, the most likely diagnosis is a primary lung cancer.

Frozen Teacher

Inspecting the chest X-ray image, the most notable finding is a well-defined mass in the right upper lobe of the lung. This mass appears as a rounded opacity without evidence of surrounding consolidation, cavitation, or pleural effusion. The mass is not associated with other typical signs of pneumonia (such as consolidation or interstitial opacities), nor does it show the typical features of a pulmonary embolism (e.g., pulmonary artery opacities, or signs of right heart strain). There is no evidence of active tuberculosis (e.g., cavitation, pleural thickening, or specific radiographic signs like hilar lymphadenopathy). The radiographic appearance is consistent with a primary lung tumor, which is most commonly a primary lung cancer. The mass is not associated with any signs of infection or inflammation. The clinical history, which includes a known diagnosis of lung cancer, further supports the diagnosis of primary lung cancer. The imaging findings are most consistent with a primary lung tumor, which is the most likely diagnosis given the presence of a right upper lobe mass and the absence of other pathologies.

Key findings:
- A clearly defined, solitary mass in the right upper lobe.
- No signs of secondary processes (e.g., pneumonia, embolism, or TB).
- The mass is not associated with pleural effusion or bronchial obstruction.
- No evidence of cavitation, which is typical of tuberculosis.
- The appearance is consistent with a primary neoplasm rather than a secondary process.

The most likely diagnosis is primary lung cancer because the imaging features—specifically a solitary mass in the right upper lobe—are highly consistent with a primary tumor. Other options are less likely:
- Pneumonia: typically shows consolidation or infiltrates, which are not seen here.
- Pulmonary embolism: would show signs of vascular obstruction, such as pulmonary artery opacities or signs of right heart strain, which are absent.
- Tuberculosis: shows specific radiographic features such as cavitation, pleural thickening, and hilar lymphadenopathy, which are not present.

Therefore, the most likely diagnosis is primary lung cancer.

T2R-R3

Inspecting the chest X-ray image, the most prominent finding is a well-defined, radiopaque mass located in the right upper lobe (RUL). Key radiographic characteristics include:
1. **Localized density**: The mass is homogenous and uniformly dense (appearing white) with no air bronchograms or cavitation—this rules out pneumonic (air-filled) or embolic (gas-filled) lesions.
2. **Well-circumscribed contour**: The mass has a sharp, defined border against the adjacent lung parenchyma, a hallmark of primary malignancy rather than infectious or inflammatory processes.
3. **No surrounding vascular or pleural changes**: There is no blunting of the costophrenic angle, no pleural effusion, no mediastinal shift, and no interstitial thickening—features absent in pneumonia (which typically shows bronchial air bronchograms and consolidation), pulmonary embolism (which shows vascular calcification or ground-glass opacities), or tuberculosis (which often presents with cavitation, hilar lymphadenopathy, and coarse reticular opacities).
4. **Absence of cavitation or cavitation-like structure**: The mass is solid and non-cavitating—consistent with a primary neoplasm, not a tuberculous lesion that may evolve into cavitation.

Now analyzing the clinical history (though not fully described in the prompt), the presence of a **right upper lobe mass** in a patient with no signs of systemic infection (no fever, no pleural effusion, no lymphadenopathy), and **no evidence of embolic signs** (e.g., wedge-shaped opacities, vascular calcifications), strongly supports a **primary lung tumor**.

Comparing this to the answer choices:
- **A: Primary lung cancer** → perfectly matches all radiographic features: well-defined, solid, non-cavitating mass in RUL with no surrounding inflammatory or vascular signs. This is the most consistent diagnosis with the imaging findings.
- **B: Pneumonia** → would show consolidation with bronchial air bronchograms (visible as hazy opacities extending into bronchi), and typically bilateral or multifocal. The absence of these features, especially the lack of air bronchograms and clear lung field margins, rules out pneumonia.
- **C: Pulmonary embolism** → would show vascular calcifications (e.g., in the pulmonary arteries), wedge-shaped opacities, or ground-glass opacities with central air bronchograms. None of these are present; the image shows only a solid, non-vascular mass.
- **D: Tuberculosis** → would typically show cavitation (even if partial), hilar lymphadenopathy (especially in the upper lobes), and coarse reticular opacities with possible pleural thickening. The absence of cavitation, lymphadenopathy, and reticulonodular opacities makes this diagnosis entirely inconsistent with the X-ray.

Therefore, the radiographic evidence—specifically the **solid, well-circumscribed, non-cavitating mass in the right upper lobe with no associated inflammatory or vascular changes**—is **uniquely consistent with primary lung cancer**. All other options are clinically and radiologically infeasible given the absence of supporting features. The only diagnosis that fully accounts for all observed findings is primary lung cancer.

Figure 12: Qualitative comparison of CoTs generated by different models.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction state the main contributions of T2R, including iterative Teacher self-competition, comparison-guided Reasoner optimization, and the case-wise reward design. The claims are supported by the experimental results and training-dynamics analyses reported in the main paper.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We discuss limitations and future work in Appendix F, including the dependence on LLM-as-a-Judge quality, the computational overhead of generating reference CoTs and pairwise comparisons, and the current use of a unified comparison prompt.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: The paper does not present formal theorems, but it includes derivations for the Reasoner objective. Detailed derivations for the sign-preserving non-degenerate case and the degenerate-case reconstruction are provided in Appendix C.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: We provide the main experimental setup in Section 4.1, including model scales, training data construction, staged data splits, baselines, evaluation benchmarks, and the answer-judging protocol. Additional reproduction details are provided in Appendix B for GRPO hyperparameters, optimizer and rollout settings, compute configuration, and prompt construction, and in Appendix A for LLM-as-a-Judge models, prompts, fallback rules, and evaluation procedures.

Guidelines:

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

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

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

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

   Answer: [Yes]

   Justification: We provide an anonymized code package in the Supplementary Material, including the Reward Service implementation, RL training code, configuration files, and detailed instructions for reproducing the main experiments. The data used in this work are publicly available with appropriate references provided in the main text, and the data construction procedure is described in detail in Section 4.1.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Section 4.1 specifies the experimental protocol, including the base models, training data, staged data splits, baseline definitions, evaluation benchmarks, answer-judge model, decoding strategy, and reported metric. Appendix B provides the detailed GRPO training setup, optimizer, rollout settings, batch sizes, sequence lengths, compute configuration, and prompt templates. Appendix A further describes the LLM-as-a-Judge settings used for CoT comparison, open-ended answer grading, and the Judge Reward baseline.

   Guidelines:

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

7. **Experiment statistical significance**

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

Answer: [No]

Justification: We do not report error bars or statistical significance tests because full multi-seed training across all model scales and baselines would be computationally expensive. To reduce variability and ensure fair comparison, all models are trained using the same random seed under matched data and optimization budgets; we additionally evaluate across six CXR open-ended VQA benchmarks and use repeated LLM-as-a-Judge evaluation with majority voting for final answer grading.

Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Appendix B reports the compute setup, including the use of two nodes with 8 NVIDIA L40 GPUs each, GPU memory, the separation between RL training and reward-service nodes, and the judge/Teacher serving configuration.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

Justification: We have reviewed the NeurIPS Code of Ethics and believe the work conforms to it. The study uses existing CXR VQA datasets and focuses on improving reasoning supervision for research purposes, without deploying a clinical decision-making system.

Guidelines:

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

10. **Broader impacts**

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

Answer: [Yes]

Justification: Appendix G provides a detailed discussion of the potential societal impacts.

Guidelines:

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

11. **Safeguards**

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

Answer: [N/A]

Justification: This work does not involve risks for misuse.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: The assets utilized in this paper are appropriately referenced and accessible to the public.

    Guidelines:

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

13. **New assets**

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

    Answer: [Yes]

    Justification: The new assets introduced by this work consist of the implementation of the T2R training pipeline, including the Reward Service, RL training code, configuration files, prompt templates, and reproduction instructions. These assets are documented and packaged in the Supplementary Material; the paper does not introduce a new dataset or benchmark.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or newly collected human-subject data.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve new human-subject experiments or interactions with human participants. It uses existing datasets and does not collect new patient data.

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [Yes]

    Justification: LLM/VLMs are central to both the proposed method and evaluation. The Reasoner and Teacher are initialized from Qwen3-VL-Instruct models and trained under the proposed T2R framework, as described in Section 3 and Section 4.1. In addition, we use LLM-as-a-Judge models for pairwise CoT comparison, open-ended answer grading, and the Judge Reward baseline; the judge models, prompts, fallback rules, and evaluation procedures are described in Appendix A.

    Guidelines:

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