Look Before You Leap: Self-Evolving Clinical Reasoning with Psychometric Preference Optimization for Radiology Report Generation

Qi Wang[1,2] **Yunfeng Min**[1] **Liwei Huang**[1] **Zeyu Zhang**[3] **Di Gai**[1] **Jun Liu**[2*]

[1]School of Mathematics and Computer Sciences, Nanchang University
[2]School of Computing and Communications, Lancaster University
[3]School of Information Science and Technology, University of Science and Technology of China

wangqi@ncu.edu.cn {yunfengmin7257,liweihuang}@email.ncu.edu.cn
zeyuzhang10@mail.ustc.edu.cn gaidi@ncu.edu.cn j.liu81@lancaster.ac.uk

Abstract

Despite the remarkable progress of large Vision-Language Models (VLMs) in Radiology Report Generation (RRG), existing black-box autoregressive generation paradigms lack intermediate clinical logical reasoning. This deficiency renders such models prone to generating linguistically fluent reports with erroneous diagnostic findings, which manifest as plausible hallucinations and thereby undermine clinical reliability. To this end, we propose a novel RRG paradigm termed "Look-Before-You-Leap", aimed at enhancing the clinical logical reasoning capabilities of the model to improve diagnostic accuracy. Specifically, we first construct a self-evolving Radiological Skill (RadSkill) repository to guide the teacher model in distilling high-confidence explicit clinical reasoning, and further internalize this reasoning capability into the implicit diagnostic intuition of the RRG baseline. Building upon this foundation, a novel preference optimization method termed Psychometric Preference Optimization (PSY-PO) is proposed. Grounded in Item Response Theory (IRT) from psychometrics, this method precisely samples clinically misleading dispreferred reports for preference optimization, thereby effectively mitigating plausible hallucinations. Extensive experimental results demonstrate that our proposed method significantly enhances both the generation quality and trustworthiness of radiology reports.

1 Introduction

Radiology Report Generation (RRG) [23][50] seeks to leverage generative AI to automatically transcribe radiological images into standardized clinical descriptions, ultimately boosting diagnostic efficiency and precision. Driven by the rapid evolution of large Vision-Language Models (VLMs) [17][39] and their broader applications [36] [7], recent research in RRG has witnessed groundbreaking progress. Nevertheless, empowering models to reliably yield clinically logical reports, especially when dealing with heterogeneous disease manifestations in chest X-ray images, poses potential challenges.

Most existing studies on RRG predominantly adopt the autoregressive generation paradigm powered by VLMs. These methods transform radiological image features into coherent clinical descriptions via supervised next-token prediction on image-report pairs. However, such end-to-end mechanisms are inherently non-transparent, rendering the underlying reasoning process uninterpretable and

*Corresponding author.

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

neglecting the logical progression intrinsic to clinical diagnosis, thus failing to replicate the rigorous cognitive processes of radiologists. To make the reasoning process more transparent, recent research attempts to explicitly construct clinical reasoning trajectories, such as medical chain-of-thought [12], by incorporating radiologist annotations or leveraging Large Language Models (LLMs) to guide the model toward explicit reasoning. Nevertheless, prohibitive manual annotation costs and the reliance on static pseudo-annotations generated via prompt engineering fall short of providing high-fidelity reasoning templates. Fundamentally, these one-off static reasoning patterns lack a dynamic evolution mechanism, preventing models from continuous self-reflection and dynamic knowledge internalization grounded in historical diagnostic experience, as radiologists do.

The black-box generation mechanism without explicit clinical reasoning constraints not only weakens the interpretability of the diagnostic process, but also further encourages the model to pursue superficial lexical similarity with ground-truth reports rather than clinical factual relevance. This training tendency makes the model prone to generating misleading hallucinated reports, thereby compromising diagnostic reliability. To mitigate plausible hallucinations, recent efforts [45] [28] have sought to incorporate human preference alignment following Supervised Fine-Tuning (SFT), compelling the model to align with the clinical diagnostic preferences of radiologists. Among these, Direct Preference Optimization (DPO) [31] has emerged as a highly efficient alternative, as it achieves preference alignment without requiring an explicit reward model. However, the efficacy of preference optimization is inherently constrained by the construction quality of the preferred and dispreferred responses. Existing construction strategies often overlook the clinical relevance between paired responses. Consequently, this induces shortcut learning, enabling the model to trivially distinguish between them through superficial textual cues. This shortcut learning phenomenon drastically diminishes the adversarial effectiveness of preference optimization, restricting the quality of generated reports in complex clinical scenarios. Additionally, current preference pair construction typically relies on single-model scoring, entangling alignment quality with the robustness of the scoring model and limiting reliable preference feedback in complex clinical contexts.

To address the above limitations, we propose a new **Look-Before-You-Leap** RRG paradigm, which reformulates trustworthy report generation as a progressive process of explicit reasoning enhancement, diagnostic capability internalization, and clinical preference constraint. In this paradigm, **"Look"** requires the model to incorporate explicit clinical logic and internalize diagnostic intuition before report generation, breaking blind end-to-end generation; **"Leap"** constrains generation through strict preference optimization to suppress plausible hallucinations. Following this design, we develop three core modules: Self-Evolving Clinical Reasoning (SECR) and Grounded Report Generation (GRG) jointly implement Look, while Psychometric Preference Optimization (PSY-PO) implements Leap. Unlike methods relying on human annotations or LLM-generated static reasoning pseudo-labels, SECR constructs a dynamically updatable RadSkill repository to guide the teacher model in generating explicit clinical reasoning trajectories. It further builds a dynamic evolution feedback loop from failed skill invocation trajectories to revise skill descriptions and diagnostic workflows, thereby transforming failure experience into reliable clinical reasoning supervision. GRG then avoids producing lengthy reasoning trajectories during inference. Instead, it injects explicit reasoning supervision into the RRG baseline and removes the intermediate reasoning chain to train direct report generation, allowing explicit reasoning capability to be internalized as implicit diagnostic intuition. Furthermore, in the post-training stage, unlike preference optimization methods that construct rejected responses through single-model scoring, random perturbation, or superficial textual differences, PSY-PO introduces Item Response Theory (IRT) [27][26] into RRG preference optimization for the first time. PSY-PO models historical model checkpoints as examinees with different diagnostic abilities, treats candidate hallucination reports as test items, and uses Fisher information to select hard negative reports that best expose the clinical blind spots of the current model, thereby constructing more discriminative preference pairs. Through these designs, Look-Before-You-Leap differs from traditional black-box generation, static reasoning supervision, and heuristic preference construction methods, while enhancing clinical reasoning before report generation and hallucination suppression during generation. In summary, our main contributions are as follows:

1. This paper proposes SECR and GRG modules to enhance clinical reasoning before report generation. SECR distills diagnostic experience from failure trajectories via a dynamic evolution loop, while GRG internalizes the resulting reasoning supervision into the RRG baseline for direct report generation. This design addresses black-box generation limitations and better simulates radiologist diagnostic reasoning.

2. We pioneer the introduction of IRT from psychometrics into preference optimization for RRG, and propose a novel PSY-PO, which effectively alleviates plausible hallucinations where model outputs appear linguistically coherent yet fundamentally violate clinical facts.
3. Extensive multi-dimensional metric evaluations on two public benchmarks (MIMIC-CXR [14] and IU X-Ray [4]) demonstrate that the proposed method exhibits significant superiority in terms of both the factual correctness and clinical trustworthiness of the generated reports.

2 Related work

2.1 Radiology Report Generation

RRG [24][34][37], a challenging task in medical artificial intelligence, aims to translate visual features into accurate and coherent clinical diagnostic reports. Early works represented by R2Gen [1] and HRGR-Agent [19], predominantly employ encoder-decoder architectures, feeding vision-encoded features into a text decoder to generate descriptions. To further enhance generation quality, several works focus on integrating cross-modal attention mechanisms [46][35], or optimizing the model architecture [2][42]. In parallel, another mainstream research line strives to incorporate prior medical knowledge to mitigate clinical factual deviations during generation. Representative explorations include medical knowledge graph integration [9][22], disease classification label guidance [13][40], and retrieval-augmented generation (RAG) [11]. With the explosive advancement of VLMs, some studies [20][25] explore RRG built upon foundation models. Despite exponential leaps in report fluency, VLMs inevitably produce plausible hallucinations due to their inherent end-to-end black-box autoregressive paradigms. This critical limitation constitutes the core motivation of our work.

2.2 Preference Optimization in Vision-language Models

To overcome the inherent limitation of SFT in driving VLMs into rigid generation paradigms [3][48][8], preference optimization has been widely integrated into the post-training phase to enhance cross-task and cross-modal generalization. Current preference optimization approaches for VLMs are broadly twofold. The first is the reward-based reinforcement learning (RL) paradigm, which aims to construct explicit reward functions via proxy reward models or verifiable heuristic rules, thereby driving the iterative alignment of the policy network (e.g., PPO [32] and GRPO [33]). The second is the reward-free direct alignment paradigm, typified by DPO [31]. This research line transforms reward modeling into a classification loss over preference pairs, elegantly bypassing RL complexities and proxy reward model training. Given their stable training and lightweight implementation, such methods [52][15] have become a mainstream alignment approach for post-training medical vision-language models. However, the preference data in existing DPO frameworks is predominantly constructed via single-model synthesis, lacking rigorous clinical relevance constraints. Consequently, its efficacy in mitigating plausible hallucinations remains highly limited. To address this, our work innovatively incorporates IRT from psychometrics, employing a collective decision-making mechanism to sample highly deceptive hard negative samples for preference pair construction, exploring an effective solution to suppress plausible hallucinations.

3 Preliminaries on Item Response Theory

IRT is a statistical framework for modeling the relationship between the latent trait of an examinee and item-level parameters. It estimates the probability of a correct response to a given item based on the interaction between the latent trait and item characteristics. In the basic one-parameter logistic 1PL model, also known as the Rasch model, each item is characterized only by a difficulty parameter. Let θ_m denote the latent trait level of examinee m and let b_k denote the difficulty parameter of item k. The probability $P_k(\theta_m)$ that examinee m correctly answers item k is defined in Eq. (1):

$$P_k(\theta_m) = \frac{1}{1 + \exp(-(\theta_m - b_k))}. \tag{1}$$

The 1PL model assumes identical discrimination across items and does not consider the effect of guessing behavior among examinees with low latent trait levels on $P_k(\theta_m)$. To address these

limitations, the three-parameter logistic 3PL model further introduces the discrimination parameter a_k and the guessing parameter c_k. In the 3PL model, $P_k(\theta_m)$ is defined in Eq. (2):

$$P_k(\theta_m) = c_k + \frac{1 - c_k}{1 + \exp(-a_k(\theta_m - b_k))}. \tag{2}$$

In Eq. (2), the characteristics of item k are jointly determined by the three parameters (a_k, b_k, c_k). Specifically, a_k denotes the discrimination parameter, which measures the ability of an item to distinguish examinees with different latent trait levels. A larger a_k indicates a steeper item response curve around the difficulty point. The parameter b_k denotes the location of an item on the latent trait scale, corresponding to item difficulty. The parameter c_k denotes the lower asymptote of the response curve, capturing the guessing probability of low-ability examinees.

4 Methodology

Figure 1 illustrates the overall pipeline of the proposed Look-Before-You-Leap paradigm, which consists of three progressive modules: SECR, grounded report generation (GRG), and PSY-PO. Given the training set $\mathcal{D} = \{(I_n, Y_n)\}_{n=1}^{N}$, we first construct a clinical reasoning generation subset $\mathcal{D}_c = \{(I_i, Y_i)\}_{i=1}^{N_c}$ through label-balanced sampling based on CheXpert disease labels. For each sample in $\mathcal{D}_c$, we employ the teacher model π_t, together with the radiology image I_i, the ground truth report Y_i, and the RadSkill repository $\mathcal{S} = \{s_k\}_{k=1}^{M}$, to generate the explicit clinical reasoning trajectory R_i^*. Meanwhile, the dynamic evolution loop uses the RRG baseline π_θ as the skill invocation agent, and continuously revises and expands the RadSkill repository $\mathcal{S}$ according to its failed skill invocation records, thereby producing more reliable clinical reasoning supervision signals. In the GRG module, the RRG baseline π_θ first fits R_i^* to learn the explicit clinical reasoning trajectory. Then, the RRG baseline π_θ removes the intermediate reasoning output and is further fine-tuned for report generation on the full training set $\mathcal{D} = \{(I_n, Y_n)\}_{n=1}^{N}$, using only I_n as input and Y_n as supervision. This enables the model to internalize explicit reasoning ability into implicit diagnostic intuition. In the PSY-PO module, we regard candidate reports as test items in IRT and treat the historical model checkpoints saved during GRG as examinees with different latent ability levels. Based on the responses of checkpoints to candidate reports, we construct preference signals and adopt the 3PL model to characterize each candidate report with discrimination a_k, difficulty b_k, and guessing c_k. By estimating these psychometric parameters, we identify hard negative reports that are fluent in surface form but highly misleading in clinical content, and then construct high-quality preference pairs. Ultimately, the model is optimized with these preference pairs to generate trustworthy reports that better align with radiologist preferences.

4.1 Self-Evolving Clinical Reasoning

SECR aims to generate reliable clinical reasoning supervision signals through explicit skill invocation and failure-driven iterative refinement. It first constructs a dynamically updatable RadSkill repository and then leverages a dynamic evolution loop to continuously refine skill definitions and reasoning workflows based on failed skill invocation trajectories, thereby improving the clinical adaptability of skill representations and the reliability of the reasoning process, as illustrated in Figure 1(a).

RadSkill Repository Construction: To generate rational clinical reasoning trajectories infused with radiological knowledge, we initialize a dynamically updatable RadSkill repository $\mathcal{S} = \{s_k\}_{k=1}^{M}$ with diagnostic priors, where M denotes the number of radiological skills. Each skill s_j consists of a description detailing its clinical diagnostic function, and a workflow defining the specific diagnostic steps to follow when invoking it. Notably, the RadSkill repository $\mathcal{S}$ covers major diagnostic dimensions in chest X-ray interpretation, including the cardiac mediastinum, pleura diaphragm, lung parenchyma, quality device, and bones soft-tissue. To simulate the prior skills required by expert radiologists during clinical reasoning, we establish a mapping between ground-truth reports and RadSkills. First, we employ RadGraph [10] to structurally parse all training set ground-truth reports to extract entity phrases containing anatomical locations and pathological observations. Subsequently, based on word frequency and skill matching analysis, we construct a two-level medical dictionary for each RadSkill. Strong matching words denote specific clinical abnormalities such as pleural effusion and pneumothorax, whereas weak matching words provide anatomical localization cues such as lung, pleura, and diaphragm. Finally, through phrase matching, if a ground-truth report contains any strong

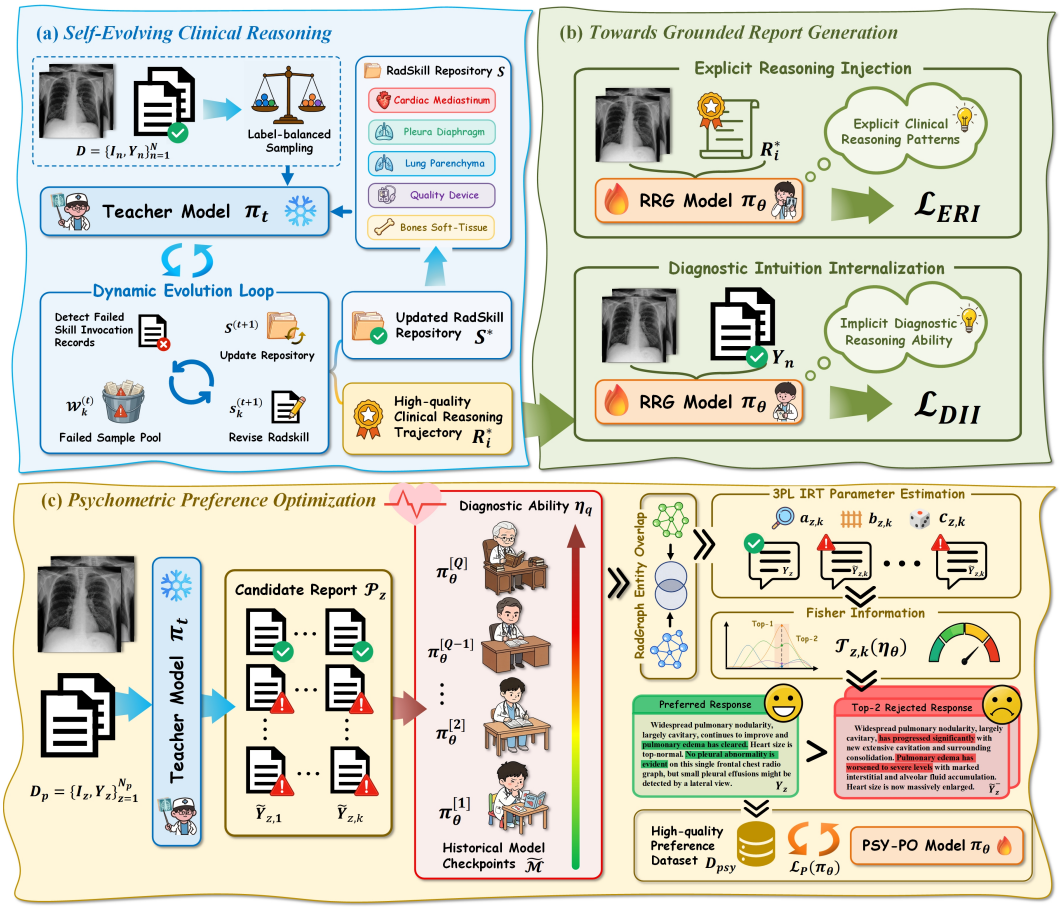


Figure 1: Pipeline of the Look-Before-You-Leap RRG paradigm. SECR first generates reliable explicit clinical reasoning trajectories via RadSkills, which are then internalized into diagnostic intuition through GRG. Ultimately, PSY-PO identifies clinically misleading hard negatives and further aligns the RRG baseline π_θ to generate trustworthy radiology reports.

matching word or at least two weak matching words, its corresponding RadSkill is designated as the prior skill for that report. Consequently, we obtain the skill supervision set $S_i^{(gt)} \subseteq \mathcal{S}$ for the i-th sample, which serves for subsequent skill invocation supervision and provides prior skill guidance for the dynamic evolution loop. **Details regarding the RadSkill repository and the two-level medical dictionary are provided in the Appendix C.2.**

Dynamic Evolution Loop: Before evolution, for any sample $(I_i, Y_i) \in \mathcal{D}_c$, we input the radiology image I_i, the ground-truth report Y_i, and the corresponding prior skill set $S_i^{(gt)}$ into the teacher model π_t to generate an initialized explicit clinical reasoning trajectory $R_i^{(0)}$ for sample i. Rather than directly serving as the final clinical reasoning supervision signal, this trajectory undergoes continuous revision during subsequent iterations alongside updates to the RadSkill repository, thereby forming a higher-quality explicit clinical reasoning supervision signal. To simulate the continuous revision process driven by failure experiences in clinical diagnosis, we design a dynamic evolution loop. Specifically, in the t-th skill-evolution epoch, the frozen RRG baseline model π_θ autonomously predicts and invokes the required skill set $S_i^{(t)}$ for sample i, conditioned on the current epoch updated reasoning trajectory $R_i^{(t)}$ and the RadSkill repository $\mathcal{S}^{(t)}$. Subsequently, we calculate the F1 score $F_i^{(t)} = 2|S_i^{(gt)} \cap S_i^{(t)}|/(|S_i^{(gt)}| + |S_i^{(t)}|)$. When $F_i^{(t)} < \tau$ (preset to 0.5), sample i is considered a failed skill invocation sample. Accordingly, the index set of failed samples in the t-th epoch is defined as $\mathcal{L}^{(t)} = \{i \mid F_i^{(t)} < \tau\}$. For any $i \in \mathcal{L}^{(t)}$, we further compute the skill error set $\Delta_i^{(t)} = S_i^{(gt)} \triangle S_i^{(t)}$, where $\triangle$ denotes the symmetric difference. This operation

characterizes missed or misinvoked skills, indicating failed sample i serves as a candidate for updating the skills within $\Delta_i^{(t)}$. Finally, for each skill $s_k^{(t)}$, we construct a corresponding failed sample pool $\mathcal{W}_k^{(t)} = \left\{ (I_i, Y_i, R_i^{(t)}) \mid i \in \mathcal{L}^{(t)},\ s_k^{(t)} \in \Delta_i^{(t)} \right\}$.

To update skill definitions from failure cases, the ground-truth reports Y_i in the failure sample pool $\mathcal{W}_k^{(t)}$ are further matched against a predefined medical lexicon to extract frequent failure cues $C_k^{(t)}$. These cues reflect missing elements in the current skill description or high-frequency pathological findings and anatomical localization information that are insufficiently covered by the workflow. Then, the current skill $s_k^{(t)}$, the skill-specific failure trajectory set $R_k^{(t)}$ constructed from 15 randomly sampled failure cases in $\mathcal{W}_k^{(t)}$, and the extracted failure cues $C_k^{(t)}$ are jointly fed into the teacher model π_t. The teacher model revises the description and workflow of $s_k^{(t)}$ to produce the updated skill $s_k^{(t+1)} = \pi_t(s_k^{(t)}, R_k^{(t)}, C_k^{(t)})$. After the t-th skill-evolution epoch, the updated RadSkill repository $\mathcal{S}^{(t+1)}$ is reused by the teacher model π_t to regenerate optimized explicit clinical reasoning trajectories $R_i^{(t+1)}$ for samples in the failure sample pool $\mathcal{W}_k^{(t)}$. These regenerated trajectories serve as input for the next skill-evolution epoch. For samples that do not trigger the failure criterion, their reasoning trajectories remain unchanged. After T skill-evolution epochs, the closed-loop iteration terminates, yielding the final high-confidence reasoning trajectories $R_i^* = R_i^{(T)}$ and the updated RadSkill repository $\mathcal{S}^* = \mathcal{S}^{(T)}$.

4.2 Towards Grounded Report Generation

To reduce the dependence of clinical reasoning trajectory generation on lengthy intermediate steps, we further internalize explicit clinical reasoning into the implicit diagnostic capability of the model, thereby supporting logically consistent direct report generation, as illustrated in Figure 1(b).

Explicit Reasoning Injection: To enable the RRG baseline to learn explicit clinical reasoning patterns, we use the high-quality reasoning trajectories R_i^* as supervised texts to perform explicit reasoning injection on π_θ. Here, R_i^* denotes the final explicit clinical reasoning trajectory retained for the i-th sample after filtering and updating through the dynamic evolution loop, which may originate from different skill-evolution epochs. Specifically, conditioned on the radiology image I_i, the RRG backbone model π_θ autoregressively fits $R_i^* = \{r_{i,1}^*; r_{i,2}^*; \ldots; r_{i,L_i}^*\}$, thereby learning the mapping from medical images to explicit clinical reasoning. This process is optimized with the standard autoregressive cross-entropy loss $\mathcal{L}_{ERI} = -\sum_{i=1}^{N_c} \sum_{l=1}^{L_i} \log P_\theta(r_{i,l}^* \mid r_{i,<l}^*, I_i)$, where N_c denotes the number of samples in $\mathcal{D}_c$, L_i denotes the token length of the reasoning trajectory for the i-th sample, $r_{i,l}^*$ denotes the l-th reasoning token, and P_θ denotes the output probability of π_θ.

Diagnostic Intuition Internalization: Although explicit clinical reasoning can enhance the ability of the model to organize image findings logically, the report generation stage should not rely on lengthy intermediate reasoning chains. Therefore, diagnostic intuition internalization removes explicit reasoning trajectories from the input and directly fits the ground-truth report Y_n conditioned only on the radiology image I_n. Through this training process, the RRG backbone model π_θ compresses the explicit reasoning ability acquired in the previous stage into an implicit diagnostic representation, enabling it to generate the final radiology report without intermediate reasoning chains. The training objective of this stage is defined as $\mathcal{L}_{DII} = -\sum_{n=1}^{N} \sum_{h=1}^{H_n} \log P_\theta(y_{n,h} \mid y_{n,<h}, I_n)$, where H_n and $y_{n,h}$ denote the report length and the h-th token of the n-th ground-truth report, respectively.

4.3 Psychometric Preference Optimization

Despite the internalization of explicit reasoning into diagnostic capability, direct report generation may still produce plausible hallucinations that are fluent on the surface but clinically incorrect. Existing methods usually mitigate hallucinations through post-SFT preference optimization, where preferred/dispreferred pairs are constructed using human-crafted rules, teacher-model scores, or text similarity. However, these methods lack a statistical measure of the clinical misleadingness of negative samples, and their preference feedback often relies on a single model or heuristic rules, making it difficult to reliably identify hard negative reports with real training value.

To this end, as shown in Figure 1(c), we introduce IRT into RRG preference optimization for the first time and propose PSY-PO, which models historical snapshots as examinees to identify clinically misleading hard negatives and construct more discriminative preference pairs. Specifically, we first sample a subset $\mathcal{D}_p = \{(I_z, Y_z)\}_{z=1}^{N_p}$ from the training set $\mathcal{D}$ for preference optimization. For the z-th sample in $\mathcal{D}_p$ and its reference report Y_z, the teacher model π_t perturbs key pathological findings, anatomical locations, or disease severity while preserving language fluency and radiological reporting style, producing a candidate hallucination pool $\mathcal{P}_z = \{\tilde{Y}_{z,k}\}_{k=1}^{K_z}$, where $\tilde{Y}_{z,k}$ denotes the k-th candidate hallucination report for the z-th sample. Meanwhile, we regard the historical model checkpoints saved during GRG training as "examinees" with different latent diagnostic abilities in IRT, denoted as $\widetilde{\mathcal{M}} = \left\{\pi_\theta^{[q]}\right\}_{q=1}^{Q}$. For the q-th checkpoint, its latent ability is obtained by applying Z-score normalization to its clinical-entity F1 score F_q, i.e., $\eta_q = (F_q - \mu_F)/(\sigma_F + \epsilon)$, where μ_F and σ_F denote the mean and standard deviation of the F1 scores across all checkpoints, and ϵ is a small constant for numerical stability.

Subsequently, each candidate hallucination report is treated as a test item in IRT. We use RadGraph to extract the medical entity sets from the candidate hallucination report $\tilde{Y}_{z,k}$ and the report $\hat{Y}_z^{[q]}$ generated by the q-th checkpoint for the same image, denoted as $\mathcal{E}(\tilde{Y}_{z,k})$ and $\mathcal{E}(\hat{Y}_z^{[q]})$, respectively. The medical entity overlap is computed as $m_{q,z,k} = \frac{\left|\mathcal{E}\left(\hat{Y}_z^{[q]}\right) \cap \mathcal{E}\left(\tilde{Y}_{z,k}\right)\right|}{\left|\mathcal{E}\left(\tilde{Y}_{z,k}\right)\right| + \epsilon}$, which measures the similarity between the report generated by the q-th checkpoint and the candidate hallucination report $\tilde{Y}_{z,k}$. A larger $m_{q,z,k}$ indicates that the checkpoint-generated report is closer to the candidate hallucination and is therefore more likely to be misled. To fit the binary-response form of the 3PL model, we set the threshold δ as the mean medical entity overlap $m_{q,z,k}$ across all candidate hallucination and checkpoint-generated reports. The overlap score is then discretized into a binary response variable $u_{q,z,k} = \mathbb{I}(m_{q,z,k} < \delta)$, where $u_{q,z,k} = 1$ indicates that the q-th checkpoint resists the k-th candidate hallucination of the z-th sample, and $u_{q,z,k} = 0$ indicates that it is misled. Based on the responses of all historical model checkpoints to the candidate hallucination reports, a binary response matrix is constructed. The 3PL model then combines this response matrix with the latent ability of each checkpoint and defines the probability that the q-th checkpoint successfully resists the candidate hallucination report $\tilde{Y}_{z,k}$ as Eq. (3):

$$P_{q,z,k} = P\left(u_{q,z,k} = 1 \mid \eta_q\right) = c_{z,k} + \frac{1 - c_{z,k}}{1 + \exp\left[-a_{z,k}\left(\eta_q - b_{z,k}\right)\right]}, \tag{3}$$

where $a_{z,k}$ denotes the clinical discrimination, which measures whether the candidate hallucination report $\widetilde{Y}_{z,k}$ can distinguish models with different diagnostic ability levels; $b_{z,k}$ denotes the clinical hallucination concealment, which characterizes the relative ability threshold required for a model to successfully resist this hallucination; and $c_{z,k}$ denotes the incidental resistance rate, which describes the lower-bound probability that a low-ability model avoids being misled by the hallucination merely due to language priors or incidental factors.

After obtaining the binary responses $u_{q,z,k}$ of all checkpoints to the candidate hallucination report, we estimate the parameters of $\widetilde{Y}_{z,k}$ by maximizing the log-likelihood under the 3PL-predicted resistance probability $P_{q,z,k}$, as formulated in Eq. (4):

$$(a_{z,k}, b_{z,k}, c_{z,k}) = \arg\max_{a,b,c} \sum_{q=1}^{Q} \left[u_{q,z,k} \log P_{q,z,k} + (1 - u_{q,z,k}) \log(1 - P_{q,z,k})\right]. \tag{4}$$

After obtaining the 3PL item parameters, the real-time ability of the current model π_θ is mapped into the same IRT ability space, denoted as η_θ. To quantify the training value of each candidate hallucination report for the current model, we introduce Fisher information into the 3PL framework. Let $P_{z,k}(\eta_\theta)$ denote the probability that the current model successfully resists the candidate hallucination report $\widetilde{Y}_{z,k}$. The Fisher information of this candidate hallucination report is defined as Eq. (5):

$$\mathcal{T}_{z,k}(\eta_\theta) = a_{z,k}^2 \cdot \frac{1 - P_{z,k}(\eta_\theta)}{P_{z,k}(\eta_\theta)} \left[\frac{P_{z,k}(\eta_\theta) - c_{z,k}}{1 - c_{z,k}}\right]^2, \tag{5}$$

where $P_{z,k}(\eta_\theta)$ denotes the probability that the current model resists the hallucination $\widetilde{Y}_{z,k}$.

Higher Fisher information indicates that the candidate hallucination report lies closer to the ability boundary of the current model and therefore can more precisely expose current clinical blind spots of the model (**See Appendix A.1 for the detailed proof**). To reduce the randomness introduced by a single hard negative and mitigate the shortcut learning risk caused by fixed surface-level cues, we no longer adopt a static negative sampling strategy. Instead, for the z-th sample, we dynamically select the Top-2 candidate hallucination reports with the highest Fisher information from the candidate pool $\mathcal{P}_z = \{\tilde{Y}_{z,k}\}_{k=1}^{K_z}$ as rejected responses, denoted as $\tilde{Y}_z^- = \{\widetilde{Y}_{z,1}^-, \widetilde{Y}_{z,2}^-\}$ (**See Appendix A.2 for the detailed proof**). Then, the reference report Y_z is used as the preferred response, while the selected two hard negative reports $\widetilde{Y}_z^-$ are used as rejected responses. Under the image condition I_z, they form a preference sample $(I_z, Y_z, \tilde{Y}_z^-)$. After constructing the high-quality preference dataset $\mathcal{D}_{psy} = \left\{(I_z, Y_z, \tilde{Y}_z^-)\right\}_{z=1}^{N_p}$, PSY-PO further optimizes the current policy model π_θ, encouraging it to assign a higher generation preference to the reference report than to the hard negative reports. Let π_{ref} denote the frozen reference model. The optimization objective is defined as Eq. (6):

$$\mathcal{L}_p(\pi_\theta) = -\mathbb{E}_{(I_z, Y_z, \tilde{Y}_z^-)\sim\mathcal{D}_{psy}}\left[\frac{1}{2}\sum_{r=1}^{2}\log\sigma\left(\beta\log\frac{\pi_\theta(Y_z \mid I_z)}{\pi_{ref}(Y_z \mid I_z)} - \beta\log\frac{\pi_\theta(\widetilde{Y}_{z,r}^- \mid I_z)}{\pi_{ref}(\widetilde{Y}_{z,r}^- \mid I_z)}\right)\right], \tag{6}$$

where $\sigma(\cdot)$ is the sigmoid function, and $\beta = 0.5$ controls the preference temperature.

5 Experiments

5.1 Experimental Settings

Implementation Details: Our experiments are implemented in PyTorch on NVIDIA RTX PRO 6000 96GB GPUs. The RRG baseline consists of a pre-trained RAD-DINO vision encoder and a Qwen2.5-0.5B-Instruct language decoder. During training, the vision encoder is frozen, while the visual projection layer and language decoder are fine-tuned. All images are resized to 336×336. In the SECR, we construct the clinical reasoning generation subset $\mathcal{D}_c = \{(I_i, Y_i)\}_{i=1}^{N_c}$ by sampling 2,069 and 3,931 image-report pairs from IU X-Ray and MIMIC-CXR, respectively, with $N_c = 6{,}000$. Qwen3.5-9B serves as the teacher model to generate explicit clinical reasoning trajectories for $\mathcal{D}_c$. The maximum number of generated tokens is 512, the batch size is 24, and explicit reasoning injection is trained for 5 epochs. During diagnostic intuition internalization, the RRG baseline is further trained for report generation for 15 epochs. In the PSY-PO, we construct the preference optimization subset $\mathcal{D}_p = \{(I_z, Y_z)\}_{z=1}^{N_p}$, where $N_p = 10{,}000$ for MIMIC-CXR and $\mathcal{D}_p$ is the full training set for IU X-Ray. Preference alignment is performed on $\mathcal{D}_p$ for 3 epochs. **The detailed prompt settings used in our framework are provided in Appendix C.1.**

Datasets and Evaluation Metrics: Our model is evaluated on two benchmarks, IU X-Ray and MIMIC-CXR. Following [1], IU X-Ray is split into training, validation, and test sets at a 7:1:2 ratio, while the official split is used for MIMIC-CXR. All experiments generate and evaluate the Findings section of radiology reports. We report two categories of metrics: natural language generation (NLG) and clinical efficacy (CE). NLG metrics include BLEU-4 (BL-4), ROUGE-L (RG-L), and METEOR (MTR) for textual similarity, while CE metrics use CheXbert to extract 14 clinical observations and compute Precision (P), Recall (R), and F1-Score (F1) for diagnostic consistency.

5.2 Ablation Study

Performance of different ablation modules. Table 1 reports the contribution of each ablation module under our Look-Before-You-Leap paradigm. The results lead to three key observations. First, compared with "+SECR w/o Evo", "+SECR" consistently improves both NLG and CE metrics on MIMIC-CXR, indicating that the dynamic evolution mechanism provides more reliable clinical reasoning supervision and strengthens the diagnostic logic of the model. Second, introducing "+PSY-PO" alone also yields clear performance gains, demonstrating that psychometric preference optimization effectively mitigates clinically misleading hallucinations. Finally, "Ours" achieves the best performance on MIMIC-CXR, validating the complementary effects of reasoning enhancement and preference constraints in generating clinically logical and reliable radiology reports.

Effect of different SFT strategies. We evaluate the impact of different supervised fine-tuning strategies on RRG. As shown in Table 2, compared with "Rep SFT", "Rea+Rep SFT" directly concatenates clinical reasoning trajectories with final reports for one-stage supervised learning. Although this strategy improves F1, it leads to clear drops in BLEU-4, METEOR, and ROUGE-L, indicating that simply concatenating reasoning texts and report texts can interfere with the model learning of report format and language distribution, thereby degrading generation quality. Furthermore, "Rep→Rea SFT", which first learns report generation and then clinical reasoning, performs worse than "Ours". This suggests that placing clinical reasoning learning after report learning cannot effectively support final report generation and may weaken the acquired report generation capability. In contrast, "Ours" achieves the best results across all NLG and CE metrics, demonstrating that the SECR training strategy more effectively internalizes self-evolving clinical reasoning supervision into the diagnostic logic of the model, leading to radiology reports with better linguistic quality and stronger clinical consistency.

Table 1: Ablation study on MIMIC-CXR. "Baseline" denotes the RRG baseline; "+SECR" and "+PSY-PO" add each corresponding module individually; and "+SECR w/o Evo " removes the dynamic evolution loop from SECR.

Modules	BL-4	MTR	RG-L	P	R	F1
Baseline	0.107	0.137	0.256	0.585	0.333	0.425
+SECR **w/o** Evo	0.126	0.158	0.289	0.621	0.441	0.516
+SECR	0.144	0.172	0.292	0.634	0.474	0.542
+PSY-PO	0.155	0.175	0.307	0.617	0.446	0.518
Ours	**0.165**	**0.181**	**0.310**	**0.636**	**0.476**	**0.544**

Table 2: Comparison of different SFT strategies for RRG on IU X-Ray. "Rep SFT", "Rea+Rep SFT", and "Rep→Rea SFT" denote report-only SFT, reasoning-report joint SFT, and report-to-reasoning SFT, respectively.

SFT Strategies	BL-4	MTR	RG-L	P	R	F1
Rep SFT	0.181	0.202	0.355	0.485	0.425	0.453
Rea+Rep SFT	0.109	0.150	0.272	0.483	0.481	0.482
Rep→Rea SFT	0.110	0.196	0.304	0.529	0.385	0.445
Ours	**0.203**	**0.227**	**0.374**	**0.548**	**0.489**	**0.517**

Effect of Rejected Response Sampling. Table 3 shows the impact of different rejected response sampling strategies on preference optimization. Although "Batch Sampling" and "Single-model Sampling" can both construct rejected responses, their performance is clearly lower than that of PSY-PO-based strategies. This indicates that relying only on intra-batch textual differences or the sampling ability of a single checkpoint is insufficient to accurately identify misleading hallucination samples. Furthermore, compared with "Single-model Sampling", PSY-PO $\left(\tilde{Y}^-_{z,1}\right)$ achieves more stable improvements on both NLG and CE metrics, with F1 increasing from 0.507 to 0.525. This suggests that the IRT-based group responses of multiple checkpoints can more effectively identify hard hallucination samples. Finally, "Ours", which uses the Top-2 clinically misleading rejected responses, achieves the best performance and further improves F1 to 0.536. These results show that multiple hard negative samples cover richer clinical misleading patterns than a single negative sample, thereby mitigating shortcut learning in preference optimization. Specifically, "Batch Sampling" selects the candidate with the lowest BLEU-1 score against the GT report within the current batch, while "Single-model Sampling" combines BLEU-1 similarity and RadGraph-based entity hallucination to select rejected responses. PSY-PO $\left(\tilde{Y}^-_{z,1}\right)$ and PSY-PO $\left(\tilde{Y}^-_{z}\right)$ denote selecting the Top-1 and Top-2 candidate hallucination reports ranked by Fisher information, respectively.

Effect of Different IRT Parameterizations. We compare the effect of different IRT parameterizations on PSY-PO. As shown in Table 4, extending the IRT parameterization from 1PL to 2PL and then to 3PL changes the estimation of Fisher information and thus affects hard negative selection, determining which candidate hallucination reports are used for preference optimization. As a result, all NLG and CE metrics consistently improve. Specifically, 1PL only estimates the difficulty parameter $b_{z,k}$ of each candidate hallucination report, ignoring its difference in distinguishing models with different ability levels. Building on this, 2PL further introduces the discrimination parameter $a_{z,k}$ and achieves better performance, indicating that candidates with stronger discriminative ability provide more effective preference optimization signals. In contrast, 3PL, which incorporates all three parameters, achieves the best results across all metrics. This suggests that finer-grained psychometric modeling can better identify hallucination samples near the current ability boundary with higher training value, thereby constructing more clinically misleading preference pairs.**More ablations, visualizations, and qualitative comparisons are provided in Appendices C.4 and C.5.**

Table 3: Comparison of rejected response sampling strategies for PSY-PO on IU X-Ray.

Sampling Strategies	BL-4	MTR	RG-L	P	R	F1
Batch Sampling	0.183	0.211	0.357	0.539	0.476	0.505
Single-model Sampling	0.203	0.202	0.366	0.534	0.483	0.507
PSY-PO $\left(\tilde{Y}^{-}_{z,1}\right)$	0.211	0.216	0.373	0.539	0.511	0.525
PSY-PO $\left(\tilde{Y}^{-}_{z}\right)$	**0.213**	**0.220**	**0.375**	**0.557**	**0.517**	**0.536**

Table 4: Effect of different IRT parameterizations on PSY-PO on IU X-Ray.

IRT Models	BL-4	MTR	RG-L	P	R	F1
1PL	0.205	0.216	0.371	0.545	0.506	0.525
2PL	0.210	0.218	0.373	0.548	0.508	0.527
3PL	**0.213**	**0.220**	**0.375**	**0.557**	**0.517**	**0.536**

5.3 Comparison with the State-of-the-art Methods

Table 5 compares the proposed method with existing mainstream RRG approaches on the MIMIC-CXR and IU X-Ray datasets, covering representative method families such as curriculum learning, dynamic knowledge-enhanced contrastive learning, relational memory, anatomy-attention alignment, and pre-trained vision-language models. The results show that Look-Before-You-Leap achieves competitive performance in both text generation quality and clinical factual consistency. For NLG metrics, our method achieves BLEU-4, METEOR, and ROUGE-L scores of 0.165, 0.181, and 0.310 on MIMIC-CXR, while maintaining competitive performance on IU X-Ray. For CE metrics, the proposed approach obtains a strong F1 score of 0.544 on MIMIC-CXR. These findings show that our method improves report quality and clinical trustworthiness through reasoning internalization and psychometric preference constraints.

Table 5: Comparison with state-of-the-art methods on MIMIC-CXR and IU X-Ray.

Methods	Encoder	Decoder	MIMIC-CXR						IU X-Ray		
			NLG Metrics			CE Metrics			NLG Metrics		
			BL-4	MTR	RG-L	P	R	F1	BL-4	MTR	RG-L
R2Gen (2020)[1]	Transformer	Transformer	0.103	0.142	0.277	0.333	0.273	0.276	0.165	0.187	0.371
CMCL (2021)[21]	CNN	HLSTM	0.097	0.133	0.281	-	-	-	0.162	0.186	0.378
METransformer (2023)[42]	ViT	Transformer	0.124	0.152	0.291	0.364	0.309	0.311	0.172	0.192	0.380
DCL (2023) [18]	ViT	Transformer	0.109	0.150	0.284	0.471	0.352	0.373	0.163	0.193	0.383
R2GenGPT (2023) [43]	Swin-Trans.	Llama2-7B	0.134	0.160	0.297	0.392	0.387	0.389	0.173	0.211	0.377
COMG+RL (2024) [6]	Transformer	Transformer	0.124	0.128	0.290	0.424	0.291	0.345	0.206	0.218	0.383
RAMT (2024) [47]	CNN	Transformer	0.113	0.153	0.284	0.380	0.342	0.355	0.165	0.195	0.377
R2Gen-Mamba (2025) [38]	Mamba	Transformer	0.110	0.141	0.284	0.483	0.325	0.389	0.176	0.208	0.382
A3Net (2025) [29]	Transformer	Transformer	0.102	0.144	0.274	-	-	-	0.175	0.199	0.381
MediVLM (2025) [5]	CLIP-ViT	GPT-2	0.123	0.149	0.293	-	-	-	0.170	0.194	0.375
AM-MRG (2026) [41]	Swin-Trans.	ClinicalBERT	0.136	0.174	0.291	0.555	0.429	0.484	0.192	0.225	0.384
DVAF (2026) [44]	VMamba	Phi-4	0.131	-	0.291	-	-	-	0.187	-	0.384
Ours	RAD-DINO	Qwen2.5-0.5B	**0.165**	**0.181**	**0.310**	**0.636**	**0.476**	**0.544**	**0.241**	**0.230**	**0.395**

6 Conclusion

In this paper, we propose a novel Look-Before-You-Leap paradigm aimed at establishing an interpretable and trustworthy RRG framework. Within this paradigm, we first introduce a Self-Evolving Clinical Reasoning module to elicit precise clinical reasoning trajectories. Building upon the internalization of such reasoning knowledge into the RRG baseline, we further substantially mitigate the generation of deceptive medical hallucinations via PSY-PO grounded in IRT. Extensive experiments on public datasets demonstrate the efficacy of our proposed method, indicating that it offers a viable solution for reliable automated clinical diagnostics.

Acknowledgments and Disclosure of Funding

This work was supported by the National Natural Science Foundation of China under Grant Nos. 62461037 and 62601354, and the Jiangxi Provincial Natural Science Foundation under Grant No. 20242BAB25078. The authors declare no competing interests.

References

[1] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.112. URL https://aclanthology.org/2020.emnlp-main.112/.

[2] Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Cross-modal memory networks for radiology report generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5904–5914, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.459. URL https://aclanthology.org/2021.acl-long.459/.

[3] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, Nicu Sebe, and Mubarak Shah. Curriculum direct preference optimization for diffusion and consistency models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2824–2834, June 2025.

[4] Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 03 2016. ISSN 1067-5027. doi: 10.1093/jamia/ocv080. URL https://doi.org/10.1093/jamia/ocv080.

[5] Debanjan Goswami, Ronast Subedi, and Shayok Chakraborty. MediVLM: A vision language model for radiology report generation from medical images. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10287–10304, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.544. URL https://aclanthology.org/2025.findings-emnlp.544/.

[6] Tiancheng Gu, Dongnan Liu, Zhiyuan Li, and Weidong Cai. Complex organ mask guided radiology report generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 7995–8004, January 2024.

[7] Yufei Guo, Yinan Wu, Haoran Duan, Guiguang Ding, and Jungong Han. Phaselora: Control-regime-conditioned low-rank adaptation for continuous-action vision-language-action policies. *arXiv preprint arXiv:2608.15285*, 2026.

[8] Jiwoo Hong, Sayak Paul, Noah Lee, Kashif Rasul, James Thorne, and Jongheon Jeong. Margin-aware preference optimization for aligning diffusion models without reference. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(6):4744–4752, Mar. 2026. doi: 10.1609/aaai.v40i6.42476. URL https://ojs.aaai.org/index.php/AAAI/article/view/42476.

[9] Zhongzhen Huang, Xiaofan Zhang, and Shaoting Zhang. Kiut: Knowledge-injected u-transformer for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19809–19818, June 2023.

[10] Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. Radgraph: Extracting clinical entities and relations from radiology reports. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran Associates, Inc., 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/c8ffe9a587b126f152ed3d89a146b445-Paper-round1.pdf.

[11] Jaehwan Jeong, Katherine Tian, Andrew Li, Sina Hartung, Subathra Adithan, Fardad Behzadi, Juan Calle, David Osayande, Michael Pohlen, and Pranav Rajpurkar. Multimodal image-text

matching improves retrieval-based chest x-ray report generation. In Ipek Oguz, Jack Noble, Xiaoxiao Li, Martin Styner, Christian Baumgartner, Mirabela Rusu, Tobias Heinmann, Despina Kontos, Bennett Landman, and Benoit Dawant, editors, *Medical Imaging with Deep Learning*, volume 227 of *Proceedings of Machine Learning Research*, pages 978–990. PMLR, 10–12 Jul 2024. URL https://proceedings.mlr.press/v227/jeong24a.html.

[12] Yue Jiang, Jiawei Chen, Dingkang Yang, Mingcheng Li, Shunli Wang, Tong Wu, Ke Li, and Lihua Zhang. Comt: Chain-of-medical-thought reduces hallucination in medical report generation. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10887699.

[13] Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. PromptMRG: Diagnosis-driven prompts for medical report generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 2607–2615, 2024. doi: 10.1609/aaai.v38i3.28038.

[14] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1): 317, December 2019. doi: 10.1038/s41597-019-0322-0. URL https://doi.org/10.1038/s41597-019-0322-0.

[15] Aishik Konwer, Zhijian Yang, Erhan Bas, Cao Xiao, Prateek Prasanna, Parminder Bhatia, and Taha Kass-Hout. Enhancing sam with efficient prompting and preference optimization for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20990–21000, June 2025.

[16] Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications, fourth edition, 2018. ISBN 9781506395661. URL https://us2.sagepub.com/en-us/nam/content-analysis/book258450.

[17] 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. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 28541–28564. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/5abcdf8ecdcacba028c6662789194572-Paper-Datasets_and_Benchmarks.pdf.

[18] Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3334–3343, June 2023.

[19] Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric Xing. Hybrid retrieval-generation reinforced agent for medical image report generation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/e07413354875be01a996dc560274708e-Paper.pdf.

[20] Chang Liu, Yuanhe Tian, Weidong Chen, Yan Song, and Yongdong Zhang. Bootstrapping large language models for radiology report generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):18635–18643, Mar. 2024. doi: 10.1609/aaai.v38i17.29826. URL https://ojs.aaai.org/index.php/AAAI/article/view/29826.

[21] Fenglin Liu, Shen Ge, and Xian Wu. Competence-based multimodal curriculum learning for medical report generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3001–3012, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.234. URL https://aclanthology.org/2021.acl-long.234/.

[22] Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13753–13762, June 2021.

[23] Kang Liu, Zhuoqi Ma, Xiaolu Kang, Yunan Li, Kun Xie, Zhicheng Jiao, and Qiguang Miao. Enhanced contrastive learning with multi-view longitudinal data for chest x-ray report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10348–10359, June 2025.

[24] Kang Liu, Zhuoqi Ma, Zikang Fang, Yunan Li, Kun Xie, and Qiguang Miao. Priorrg: Prior-guided contrastive pre-training and coarse-to-fine decoding for chest x-ray report generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(9):7206–7214, Mar. 2026. doi: 10.1609/aaai.v40i9.37657. URL https://ojs.aaai.org/index.php/AAAI/article/view/37657.

[25] Tengfei Liu, Jiapu Wang, Yongli Hu, Mingjie Li, Junfei Yi, Xiaojun Chang, Junbin Gao, and Baocai Yin. Hc-llm: Historical-constrained large language models for radiology report generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(6):5595–5603, Apr. 2025. doi: 10.1609/aaai.v39i6.32596. URL https://ojs.aaai.org/index.php/AAAI/article/view/32596.

[26] Frederic M. Lord. *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1980.

[27] Frederic M. Lord and Melvin R. Novick. *Statistical theories of mental test scores*. Addison-Wesley, 1968.

[28] Zhenjie Luo, Libin Yang, Hui Zhang, Quan Tang, Shirui Pan, Dehong Gao, and Xiaoyan Cai. Spatio -temporal-aware preference optimization for personalized radiology report generation. *Expert Systems with Applications*, 322:132364, 2026. ISSN 0957-4174. doi: 10.1016/j.eswa.2026.132364. URL https://www.sciencedirect.com/science/article/pii/S0957417426012777.

[29] Quang Vinh Nguyen, Minh Duc Nguyen, Thanh Hoang Son Vo, Hyung-Jeong Yang, and Soo-Hyung Kim. Anatomical attention alignment representation for radiology report generation. In *2025 IEEE International Conference on Image Processing (ICIP)*, pages 2826–2831, 2025. doi: 10.1109/ICIP55913.2025.11084659.

[30] Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, and Jean-Benoit Delbrouck. GREEN: Generative radiology report evaluation and error notation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 374–390, Miami, Florida, USA, nov 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.21. URL https://aclanthology.org/2024.findings-emnlp.21/.

[31] 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. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a85b405ed65c6477a4fe8302b5e06ce7-Paper-Conference.pdf.

[32] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, abs/1707.06347, 2017. doi: 10.48550/arXiv.1707.06347. URL https://arxiv.org/abs/1707.06347.

[33] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, abs/2402.03300, 2024. doi: 10.48550/arXiv.2402.03300. URL https://arxiv.org/abs/2402.03300.

[34] Shanshan Song, Hui Tang, Honglong Yang, and Xiaomeng Li. Ddatr: Dynamic difference-aware temporal residual network for longitudinal radiology report generation. *IEEE Transactions on Medical Imaging*, 44(12):5345–5357, 2025. doi: 10.1109/TMI.2025.3591364.

[35] Xiao Song, Xiaodan Zhang, Junzhong Ji, Ying Liu, and Pengxu Wei. Cross-modal contrastive attention model for medical report generation. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2388–2397, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.210/.

[36] Xinshuai Song, Weixing Chen, Yang Liu, Weikai Chen, Guanbin Li, and Liang Lin. Towards long-horizon vision-language navigation: Platform, benchmark and method. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12078–12088, 2025. doi: 10.1109/CVPR52734.2025.01128.

[37] Zhaoli Su, Yucong Lin, Hong Song, Ruoyi Jian, Bowen Liu, and Jian Yang. Vtag: Visual-textual association guided radiology reports generation. *IEEE Transactions on Image Processing*, 34: 7391–7406, 2025. doi: 10.1109/TIP.2025.3623915.

[38] Yongheng Sun, Yueh Z. Lee, Genevieve A. Woodard, Hongtu Zhu, Chunfeng Lian, and Mingxia Liu. R2gen-mamba: A selective state space model for radiology report generation. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pages 1–4, 2025. doi: 10.1109/ISBI60581.2025.10980814.

[39] Omkar Chakradhar Thawakar, Abdelrahman M. Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Khan. XrayGPT: Chest radiographs summarization using large medical vision-language models. In Dina Demner-Fushman, Sophia Ananiadou, Makoto Miwa, Kirk Roberts, and Junichi Tsujii, editors, *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 440–448, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.bionlp-1.35. URL https://aclanthology.org/2024.bionlp-1.35/.

[40] Jun Wang, Abhir Bhalerao, Terry Yin, Simon See, and Yulan He. Camanet: Class activation map guided attention network for radiology report generation. *IEEE Journal of Biomedical and Health Informatics*, 28(4):2199–2210, 2024. doi: 10.1109/JBHI.2024.3354712.

[41] Xiao Wang, Fuling Wang, Haowen Wang, Bo Jiang, Chuanfu Li, Yaowei Wang, Yonghong Tian, and Jin Tang. Activating associative disease-aware vision token memory for llm-based x-ray report generation. *IEEE Transactions on Medical Imaging*, 45(2):583–595, 2026. doi: 10.1109/TMI.2025.3603416.

[42] Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11558–11567, June 2023.

[43] Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. R2gengpt: Radiology report generation with frozen llms. *Meta-Radiology*, 1(3):100033, 2023. ISSN 2950-1628. doi: 10.1016/j.metrad.2023.100033. URL https://www.sciencedirect.com/science/article/pii/S2950162823000334.

[44] Puzhen Wu, Hexin Dong, Yi Lin, Yihao Ding, and Yifan Peng. A disease-aware dual-stage framework for chest x-ray report generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(40):33953–33961, Mar. 2026. doi: 10.1609/aaai.v40i40.40688. URL https://ojs.aaai.org/index.php/AAAI/article/view/40688.

[45] Ting Xiao, Lei Shi, Peng Liu, Zhe Wang, and Chenjia Bai. Radiology report generation via multi-objective preference optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(8):8664–8672, Apr. 2025. doi: 10.1609/aaai.v39i8.32936. URL https://ojs.aaai.org/index.php/AAAI/article/view/32936.

[46] Di You, Fenglin Liu, Shen Ge, Xiaoxia Xie, Jing Zhang, and Xian Wu. Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 72–82, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87199-4.

[47] Ke Zhang, Hanliang Jiang, Jian Zhang, Qingming Huang, Jianping Fan, Jun Yu, and Weidong Han. Semi-supervised medical report generation via graph-guided hybrid feature consistency. *IEEE Transactions on Multimedia*, 26:904–915, 2024. doi: 10.1109/TMM.2023.3273390.

[48] Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander G Hauptmann, Yonatan Bisk, and Yiming Yang. Direct preference optimization of video large multimodal models from language model reward. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 694–717, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.30. URL https://aclanthology.org/2025.naacl-long.30/.

[49] Yabin Zhang, Chong Wang, Yunhe Gao, Jiaming Liu, Maya Varma, Justin Xu, Sophie Ostmeier, Jin Long, Sergios Gatidis, Seena Dehkharghani, Arne Michalson, Eun Kyoung Hong, Christian Bluethgen, Haiwei Henry Guo, Alexander Victor Ortiz, Stephan Altmayer, Sandhya Bodapati, Joseph David Janizek, Ken Chang, Jean-Benoit Delbrouck, Akshay S. Chaudhari, and Curtis P. Langlotz. A reasoning-enabled vision-language foundation model for chest x-ray interpretation, 2026. URL https://arxiv.org/abs/2604.00493. arXiv preprint.

[50] Qin Zhou, Guoyan Liang, Xindi Li, Jingyuan Chen, Zhe Wang, Chang Yao, and Sai Wu. Learnable retrieval enhanced visual-text alignment and fusion for radiology report generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22529–22538, October 2025.

[51] Ciyou Zhu, Richard H. Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software*, 23(4):550–560, 1997. doi: 10.1145/279232.279236.

[52] Kangyu Zhu, Peng Xia, Yun Li, Hongtu Zhu, Sheng Wang, and Huaxiu Yao. MMedPO: Aligning medical vision-language models with clinical-aware multimodal preference optimization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 80207–80222. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhu25v.html.

Technical appendices and supplementary material

A Proofs of the Main Theoretical Results **17**

A.1 Proof of Proposition 1 17

A.2 Proof of Corollary 1 18

B Training Algorithm Pseudocode **18**

C Experimental Details **19**

C.1 Prompt Template Details 19

C.2 RadSkill Repository and Medical Dictionary 23

C.3 PSY-PO Implementation Details 25

C.4 Further Ablation Results 25

C.5 Further Qualitative Results 27

C.6 Robustness and Sensitivity Analysis 31

C.7 Clinical Reliability Evaluation 32

D Case Study **34**

D.1 Examples on Explicit Clinical Reasoning 34

D.2 Examples on Final Radiology Report Generation 35

E Limitations and Future Work **35**

A Proofs of the Main Theoretical Results

A.1 Proof of Proposition 1

Proposition 1 (Fisher Information-Guided Hard Negative Sample Selection Maximizes Local Diagnostic Discriminative Information)**.** *Let the current model capability be η_θ. For the candidate hallucination report $\tilde{Y}_{z,k}$ of the z-th sample, the binary response follows the 3PL model as defined in the main text, with the successful resistance probability denoted by $P_{z,k}(\eta)$. The Fisher information of this candidate hallucination report at the current capability point η_θ is defined as:*

$$\mathcal{T}_{z,k}(\eta_\theta) = \frac{(P'_{z,k}(\eta_\theta))^2}{P_{z,k}(\eta_\theta)(1 - P_{z,k}(\eta_\theta))}. \tag{A.1}$$

Given a sufficiently small local perturbation Δ applied to the current model capability, the following Kullback-Leibler (KL) divergence approximation holds:

$$D_{\mathrm{KL}}\left(\mathrm{Bern}(P_{z,k}(\eta_\theta + \Delta)) \parallel \mathrm{Bern}(P_{z,k}(\eta_\theta))\right) = \frac{1}{2}\mathcal{T}_{z,k}(\eta_\theta)\Delta^2 + o(\Delta^2). \tag{A.2}$$

Therefore, to second order in the local perturbation Δ, selecting the candidate hallucination report with the maximum Fisher information is equivalent to selecting the hard negative sample that maximizes the variation in the response distribution near the current model capability. The resulting selection rule is formulated as follows:

$$\tilde{Y}^{\star}_{z,k} = \arg\max_{\tilde{Y}_{z,k} \in \mathcal{P}_z} \mathcal{T}_{z,k}(\eta_\theta). \tag{A.3}$$

Proof. For brevity, denote $p = P_{z,k}(\eta_\theta)$ and $p_\Delta = P_{z,k}(\eta_\theta + \Delta)$.

By the 3PL parameterization defined in the main text, $P_{z,k}(\eta)$ is continuously differentiable and satisfies $P_{z,k}(\eta_\theta) \in (0, 1)$. Therefore, applying the first-order Taylor expansion around η_θ yields:

$$p_\Delta = p + P'_{z,k}(\eta_\theta)\Delta + o(\Delta). \tag{A.4}$$

Let $\delta = p_\Delta - p$, then we have:

$$\delta = P'_{z,k}(\eta_\theta)\Delta + o(\Delta). \tag{A.5}$$

The KL divergence between the two Bernoulli distributions $\mathrm{Bern}(p_\Delta)$ and $\mathrm{Bern}(p)$ is defined as:

$$D_{\mathrm{KL}}(\mathrm{Bern}(p_\Delta) \parallel \mathrm{Bern}(p)) = p_\Delta \log \frac{p_\Delta}{p} + (1 - p_\Delta) \log \frac{1 - p_\Delta}{1 - p}. \tag{A.6}$$

Substituting $p_\Delta = p + \delta$ into the above equation, we obtain:

$$D_{\mathrm{KL}} = (p + \delta) \log \left(1 + \frac{\delta}{p}\right) + (1 - p - \delta) \log \left(1 - \frac{\delta}{1 - p}\right). \tag{A.7}$$

Utilizing the second-order Taylor expansion for the logarithmic function at $\delta = 0$, i.e., $\log(1 + x) = x - \frac{x^2}{2} + o(x^2)$, we can expand the logarithmic terms as follows:

$$\log \left(1 + \frac{\delta}{p}\right) = \frac{\delta}{p} - \frac{\delta^2}{2p^2} + o(\delta^2), \tag{A.8}$$

$$\log \left(1 - \frac{\delta}{1 - p}\right) = -\frac{\delta}{1 - p} - \frac{\delta^2}{2(1 - p)^2} + o(\delta^2). \tag{A.9}$$

Substituting these expansions back and rearranging the terms yields:

$$(p + \delta) \log \left(1 + \frac{\delta}{p}\right) = \delta + \frac{\delta^2}{2p} + o(\delta^2), \tag{A.10}$$

$$(1 - p - \delta) \log \left(1 - \frac{\delta}{1 - p}\right) = -\delta + \frac{\delta^2}{2(1 - p)} + o(\delta^2). \tag{A.11}$$

Adding the two equations, the first-order term δ cancels out, resulting in:

$$D_{\mathrm{KL}}(\mathrm{Bern}(p_\Delta) \parallel \mathrm{Bern}(p)) = \frac{\delta^2}{2}\left(\frac{1}{p} + \frac{1}{1-p}\right) + o(\delta^2) = \frac{\delta^2}{2p(1-p)} + o(\delta^2). \tag{A.12}$$

Substituting $\delta = P'_{z,k}(\eta_\theta)\Delta + o(\Delta)$ into the equation, we get:

$$D_{\mathrm{KL}}\left(\mathrm{Bern}(P_{z,k}(\eta_\theta + \Delta)) \parallel \mathrm{Bern}(P_{z,k}(\eta_\theta))\right) = \frac{(P'_{z,k}(\eta_\theta))^2}{2P_{z,k}(\eta_\theta)(1 - P_{z,k}(\eta_\theta))}\Delta^2 + o(\Delta^2). \tag{A.13}$$

According to the definition of Fisher information $\mathcal{T}_{z,k}(\eta_\theta) = \frac{(P'_{z,k}(\eta_\theta))^2}{P_{z,k}(\eta_\theta)(1-P_{z,k}(\eta_\theta))}$, the local KL divergence can be rewritten as:

$$D_{\mathrm{KL}}\left(\mathrm{Bern}(P_{z,k}(\eta_\theta + \Delta)) \parallel \mathrm{Bern}(P_{z,k}(\eta_\theta))\right) = \frac{1}{2}\mathcal{T}_{z,k}(\eta_\theta)\Delta^2 + o(\Delta^2). \tag{A.14}$$

In summary, when the local perturbation Δ is fixed and sufficiently small, maximizing this local KL divergence is equivalent to second order in the local perturbation to maximizing the Fisher information $\mathcal{T}_{z,k}(\eta_\theta)$. Therefore, selecting the candidate report with the maximum Fisher information from the hallucination pool maximizes the second-order approximation of the local KL divergence in the neighborhood of the current model capability point, thereby maximizing the local diagnostic discriminative information. That is:

$$\tilde{Y}^\star_{z,k} = \arg\max_{\tilde{Y}_{z,k} \in \mathcal{P}_z} \mathcal{T}_{z,k}(\eta_\theta). \tag{A.15}$$

□

A.2 Proof of Corollary 1

Corollary 1 (Top-k Fisher Selection Maximizes Local Diagnostic Discriminative Information)**.** *Assuming the local training information of multiple rejected responses is additive with respect to Fisher information, for any candidate subset $\mathcal{A}_z \subseteq \mathcal{P}_z$ of size k, its total local diagnostic discriminative information can be expressed as:*

$$\mathcal{T}(\mathcal{A}_z; \eta_\theta) = \sum_{\tilde{Y}_{z,\ell} \in \mathcal{A}_z} \mathcal{T}_{z,\ell}(\eta_\theta). \tag{A.16}$$

Therefore, the optimization problem of maximizing the total information of these k negative samples can be formulated as:

$$\mathcal{A}^\star_z = \arg\max_{\substack{\mathcal{A}_z \subseteq \mathcal{P}_z \\ |\mathcal{A}_z| = k}} \sum_{\tilde{Y}_{z,\ell} \in \mathcal{A}_z} \mathcal{T}_{z,\ell}(\eta_\theta). \tag{A.17}$$

Since the objective is modular and the subset cardinality is fixed, the optimal solution is obtained by selecting the k candidate hallucinated reports with the highest individual Fisher information values:

$$\mathcal{A}^\star_z = \text{Top-}k_{\tilde{Y}_{z,\ell} \in \mathcal{P}_z} \mathcal{T}_{z,\ell}(\eta_\theta). \tag{A.18}$$

Consequently, under the assumption of additive local information, selecting the Top-k candidate hallucinated reports with the highest Fisher information as rejected responses ensures the maximization of the total local diagnostic discriminative information in the neighborhood of the current model capability among all combinations of k negative samples.

B Training Algorithm Pseudocode

We further provide the overall training pipeline of our method. As shown in Algorithm 1, Look-Before-You-Leap consists of three core modules: SECR, Grounded Report Generation, and PSY-PO.

Algorithm 1 Overall Training Pipeline

Require: Dataset $\mathcal{D}_c$, teacher π_t, RRG backbone π_θ, checkpoints $\widetilde{\mathcal{M}}$, preference subset $\mathcal{D}_p$.
Ensure: Optimized π_θ, updated RadSkill repository $\mathcal{S}^*$.

▷ ***Phase 1: Self-Evolving Clinical Reasoning***
Initialize RadSkill repository $\mathcal{S}^{(0)}$ and prior skill sets $S_i^{(gt)}$
Generate initial reasoning trajectories $R_i^{(0)}$ with teacher π_t
for $t = 0$ **to** $T-1$ **do**
    Infer invoked $S_i^{(t)} \leftarrow \pi_\theta(R_i^{(t)}, \mathcal{S}^{(t)})$ and compute $F_i^{(t)}$
    Identify failures $\mathcal{L}^{(t)} = \{i \mid F_i^{(t)} < \tau\}$ and errors $\Delta_i^{(t)}$
    Extract cues $C_k^{(t)}$ from pool $\mathcal{W}_k^{(t)}$
    Update skills $s_k^{(t+1)} \leftarrow \pi_t(s_k^{(t)}, R_k^{(t)}, C_k^{(t)})$
    Regenerate $R_i^{(t+1)} \leftarrow \pi_t(I_i, Y_i, S_i^{(gt)})$ for $i \in \mathcal{L}^{(t)}$
end for
Obtain final $R_i^* = R_i^{(T)}$ and repository $\mathcal{S}^* = \mathcal{S}^{(T)}$
▷ ***Phase 2: Grounded Report Generation***
Sequentially optimize π_θ with $\mathcal{L}_{ERI}$ on R_i^* and $\mathcal{L}_{DII}$ on Y_i
Save checkpoints as $\widetilde{\mathcal{M}} = \{\pi_\theta^{[q]}\}_{q=1}^{Q}$
Set $\eta_\theta \leftarrow \mathrm{ZScore}(F_\theta)$
▷ ***Phase 3: Psychometric Preference Optimization***
Initialize $\mathcal{D}_{psy}$ and frozen reference $\pi_{ref} \leftarrow \pi_\theta$
for each sample $(I_z, Y_z) \in \mathcal{D}_p$ **do**
    Generate hallucination pool $\mathcal{P}_z = \{\tilde{Y}_{z,k}\}_{k=1}^{K_z}$ with π_t
    for each candidate $\tilde{Y}_{z,k} \in \mathcal{P}_z$ **do**
        Compute responses $u_{q,z,k} = \mathbb{I}(m_{q,z,k} < \delta)$ over $\widetilde{\mathcal{M}}$
        Estimate 3PL parameters $(a_{z,k}, b_{z,k}, c_{z,k})$ by maximizing Eq. (4)
        Compute Fisher information $\mathcal{T}_{z,k}(\eta_\theta)$ via Eq. (5)
    end for
    Select Top-2 candidates by $\mathcal{T}_{z,k}(\eta_\theta)$ as $\tilde{Y}_z^- = \{\tilde{Y}_{z,1}^-, \tilde{Y}_{z,2}^-\}$
    $\mathcal{D}_{psy} \leftarrow \mathcal{D}_{psy} \cup \{(I_z, Y_z, \tilde{Y}_z^-)\}$
end for
Optimize π_θ on $\mathcal{D}_{psy}$ by minimizing $\mathcal{L}_p(\pi_\theta)$ in Eq. (6)
return Optimized π_θ and repository $\mathcal{S}^*$

C Experimental Details

C.1 Prompt Template Details

This section summarizes the prompt templates used in Look-Before-You-Leap and explains their correspondence to the three core modules: Self-Evolving Clinical Reasoning (SECR), Grounded Report Generation (GRG), and Psychometric Preference Optimization (PSY-PO).

Prompt Design for SECR.

SECR uses three prompts to support explicit clinical reasoning generation and dynamic RadSkill evolution. First, the **Clinical Reasoning Generation Prompt** guides the teacher model to generate or regenerate explicit clinical reasoning trajectories conditioned on the medical image, reference report, and current RadSkill repository, as shown in Figure C.1. Second, the **Skill Prediction Prompt** guides the frozen RRG baseline to predict the invoked medical skill set from the current reasoning trajectory, which is used to identify failed skill invocation samples, as shown in Figure C.2. Finally, the **Dynamic Evolution Prompt** is applied after failure detection, where the target skill, failure sample pool, and extracted failure cues are provided to the teacher model to revise the corresponding RadSkill description and workflow, as shown in Figure C.3. Together, these prompts support SECR in iteratively transforming failed skill invocation trajectories into more reliable clinical reasoning supervision signals.

Prompt: Clinical Reasoning Generation

System Prompt:

You are an expert radiology reasoning assistant. Your task is to generate a concise, objective, step-by-step clinical reasoning grounded in the provided chest X-ray image, clinical report, and expert analytical skills.

Rules:
1.The content inside <*thinking*> must not exceed 120 words.
2.The clinical reasoning must be logical, clinically relevant, and concise.
3.Do not include conversational filler, disclaimers, or introductory text.
4.Do not output anything outside a single <*thinking*>...</*thinking*> block.

User Prompt:

Task: Use the image, the Clinical Report, and the Expert Analytical Skills to produce a compact clinical reasoning. Start directly with <*thinking*>.

Clinical Report: {*report*}
Expert Analytical Skills: {*target_skill*}

Output Example:
<*thinking*>
First, device check shows the endotracheal tube is suboptimally low (<2 cm from carina) needing withdrawal; other lines are adequately placed. Second, parenchymal evaluation reveals left lower lobe collapse, interstitial edema, and worsening bilateral perihilar opacities. Third, checking the framework, multiple left rib fractures exist but don't explain the central opacities. Finally, synthesizing the differential: the perihilar pattern and edema strongly suggest fluid overload (pulmonary edema). However, progressive opacities in an intubated patient mandate considering infection (ventilator-associated pneumonia). Contusion is less typical for this distribution. Conclusion: Primary impression is pulmonary edema versus infection requiring clinical correlation, plus an actionable alert for ETT adjustment.
</*thinking*>

Figure C.1: Prompt template for clinical reasoning generation.

Prompt: Skill Prediction

System Prompt:

You are an expert in medical skill selection.

User Prompt:

Task: Based on the provided Clinical Reasoning, select {*target_skill_count*}relevant skills in the Skill Bank that are explicitly supported or clearly required by the Clinical Reasoning.

Clinical Reasoning:
{*clinical_reasoning*}

Skill Bank:
{*skill_bank*}

Requirements:
1.Every selected skill must come from the Skill Bank.
2.Do not include explanations, notes, markdown, or extra text.
3.You MUST output exactly {*target_skill_count*} skill ID(s).

Output Format:
["<Skill_ID_1>", "<Skill_ID_2>", ...]

Figure C.2: Prompt template for skill prediction.

Prompt: Skill Update

System Prompt:

You are an expert chest X-ray educator refining exactly one medical reasoning skill.

Your task is to revise the target skill based on two provided inputs:
- Auxiliary Information (high-frequency keywords from the clinical reports).
- Representative Cases (expected clinical reasoning paired with chest X-ray images).

Hard constraints:
1. Preserve the original skill boundary.
2. Update the description (if necessary) and the full workflow to reflect the corrections.
3. The final updated workflow **MUST NOT** exceed 5 steps in total.
4. If integrating missing coverage would cause the workflow to exceed 5 steps, you are strictly forbidden from appending new steps. Instead, you must merge, replace, or revise existing steps to absorb the new logic.
5. Return exactly one valid JSON object and nothing else.

User Prompt:

TARGET SKILL
Skill ID: {*skill_id*}

CURRENT SKILL DEFINITION
Description: {*description*}
Workflow: {*workflow*}

AUXILIARY INFORMATION
High-Frequency Keywords: {*dynamic_cues*}
Usage Instructions: These keywords highlight standard clinical findings from cases where the current student model underperforms. Treat them as directional cues to deduce missing observation steps or logical checks.

REPRESENTATIVE CASES:
Case {*idx*}
Expected Clinical Reasoning: {*clinical_reasoning*}
[*Attached Image*]

Task: Please update the target skill based on the provided clinical cases. Return **STRICT JSON** only, without markdown code fences. Briefly explain the identified missing logic. Use this exact schema:
{
"updated_description": "Revised description of the skill.",
"updated_workflow": "The complete, revised workflow string containing the specific observation steps and logical checks."
}

Figure C.3: Prompt template for dynamic evolution.

Prompt Design for GRG.

GRG consists of two training steps and accordingly uses two prompts. In the explicit reasoning injection step, the **Explicit Reasoning Injection Prompt** guides the RRG baseline to fit the dynamically updated clinical reasoning trajectories from medical images, enabling the model to learn the mapping from visual evidence to explicit clinical reasoning processes, as shown in Figure C.4. In the diagnostic intuition internalization step, the **Diagnostic Intuition Internalization Prompt** removes intermediate reasoning chains and directly guides the model to generate final radiology reports from medical images, thereby compressing explicit reasoning capability into implicit diagnostic capability, as shown in Figure C.5.

Prompt: Explicit Reasoning Injection

System Prompt:

You are an expert radiologist.

User Prompt:

Examine this chest X-ray and generate a step-by-step reasoning process.

Figure C.4: Prompt template for explicit reasoning injection.

Prompt: Diagnostic Intuition Internalization

System Prompt:

You are a radiologist.

User Prompt:

Indication: {*indication*}. History: {*history*}. Based on the provided history, indication, and images, write a detailed and professional radiological findings report.

Figure C.5: Prompt template for diagnostic intuition internalization.

Prompt Design for PSY-PO.

PSY-PO employs the **Hallucination Injection Prompt** to construct candidate dispreferred reports for preference optimization. Specifically, the prompt guides the teacher model to inject clinically significant inaccuracies into key lesion descriptions of ground-truth reports while preserving their linguistic fluency and reporting style. Two in-context examples illustrate the hallucination injection process, and the teacher model is instructed to output only the hallucination-injected report. Multiple independently sampled responses form a candidate hallucination pool for each ground-truth report, from which IRT-based psychometric evaluation selects informative hard negatives for subsequent preference optimization, as shown in Figure C.6.

Prompt: Hallucination Injection

User Prompt:

You are provided with a ground truth report: The heart size and pulmonary vascularity appear within normal limits. A large hiatal hernia is noted. The lungs are free of focal airspace disease. No pneumothorax or pleural effusion is seen. Degenerative changes are present in the spine.. Please focus on the critical lesion descriptions within the report and inject hallucinations into these key descriptions to construct a dispreferred response. Ensure that the dispreferred responses contain significant medical inaccuracies. You may refer to the following examples for guidance:

<example 1>
Ground truth report:
Frontal and lateral radiographs of the chest redemonstrate a round calcified pulmonary nodule in the posterior right lung base, unchanged from multiple priors and consistent with prior granulomatous disease. A known enlarged right hilar lymph node seen on CT of ___ likely accounts for the increased opacity at the right hilum. A known right mediastinal lymph node conglomerate accounts for the fullness at the right paratracheal region. No pleural effusion, pneumothorax or focal consolidation is present. The patient is status post median sternotomy and CABG with wires intact. The cardiac silhouette is normal in size. The mediastinal and hilar contours are unchanged from the preceding radiograph.

Hallucination-injected response:
Frontal and lateral radiographs of the chest redemonstrate an irregular right basilar pulmonary nodule, increased in size from multiple priors and concerning for malignancy. Increased opacity at the right hilum and fullness of the right paratracheal region appear more prominent, suggesting progressive hilar and mediastinal adenopathy. Small right pleural effusion and mild adjacent right basilar airspace opacity are present. The patient is status post median sternotomy and CABG. The cardiac silhouette is mildly enlarged.
</example 1>

<example 2>
Ground truth report:
Left PICC tip is seen terminating in the region of the distal left brachiocephalic vein. Tracheostomy tube is in unchanged standard position. The heart is moderately enlarged. Marked calcification of the aortic knob is again present. Mild pulmonary vascular congestion is similar. Bibasilar streaky airspace opacities are minimally improved. Previously noted left pleural effusion appears to have resolved. No pneumothorax is identified. Percutaneous gastrostomy tube is seen in the left upper quadrant.

Hallucination-injected response:
Left PICC tip is seen terminating in the distal left brachiocephalic vein and may be slightly retracted. Tracheostomy tube remains in place. The heart is enlarged. Mild pulmonary vascular congestion has increased, and bibasilar patchy airspace opacities have worsened, which may reflect edema or superimposed infection. Small left pleural effusion persists. A trace left apical pneumothorax is present. Percutaneous gastrostomy tube projects over the left upper quadrant.
</example 2>

Your response should consist only of the hallucination-injected response.

Figure C.6: Prompt template for hallucination injection.

C.2 RadSkill Repository and Medical Dictionary

RadSkill Repository and Matching Dictionary.

To ensure the executability of the prompts in clinical reasoning generation and skill evolution, we also provide the initialized RadSkill repository. This repository covers the major diagnostic dimensions in chest X-ray interpretation, including the cardiac mediastinum, pleural diaphragm, lung parenchyma, quality device, and bones soft-tissue. Each RadSkill consists of a description and a workflow, and the initialized repository is shown in Figure C.7.

To further support automatic matching between ground-truth reports and prior RadSkills, we assign a two-level medical dictionary to each RadSkill. Specifically, strong matching keywords correspond to specific clinical abnormalities, while weak matching keywords provide anatomical localization cues. The complete matching dictionary is shown in Figure C.8.

RADSKILL Repository

Skill 1: Cardiac_Mediastinum



Description:
Evaluate the size and contour of the cardiac silhouette and the configuration of the mediastinum.

Workflow:
1. Assess the cardiothoracic ratio to determine cardiomegaly.
2. Assess the definition of the right and left heart borders and check for mediastinal widening.
3. Evaluate the aortic knob for tortuosity or calcification.

Skill 2: Pleural_Diaphragm



Description:
Evaluate the pleural spaces and diaphragm for the presence of fluid or free air.

Workflow:
1. Examine the costophrenic angles for blunting, suggesting pleural effusion.
2. Assess the lung apices for pleural lines indicating pneumothorax.
3. Evaluate the contour and position of the hemidiaphragms.

Skill 3: Lung_Parenchyma



Description:
Evaluate the aeration of the lung parenchyma bilaterally, assessing for airspace or interstitial abnormalities.

Workflow:
1. Divide the lungs into upper, mid, and lower zones and perform side-to-side comparisons.
2. Assess for opacities, consolidation, atelectasis, nodules, or other abnormalities.
3. Characterize the distribution and morphology of findings (e.g., peribronchial, diffuse, focal) to suggest underlying processes such as infection or edema.

Skill 4: Quality_Device



Description:
Assess the presence, integrity, and positioning of implanted medical devices or foreign bodies.

Workflow:
1. Survey the thorax for implanted devices (e.g., pacemakers, orthopedic hardware) or lines and tubes.
2. Verify that the distal tips of endotracheal tubes (ETT), nasogastric tubes, and central venous catheters (CVC) terminate in appropriate anatomical positions.

Skill 5: Bones_SoftTissue



Description:
Evaluate the osseous structures of the thorax and the surrounding soft tissues.

Workflow:
1. Trace the clavicles and ribs to assess cortical integrity and identify fractures.
2. Evaluate the thoracic spine, shoulder joints, and soft tissues for abnormalities.

Figure C.7: Initialized RadSkill repository.

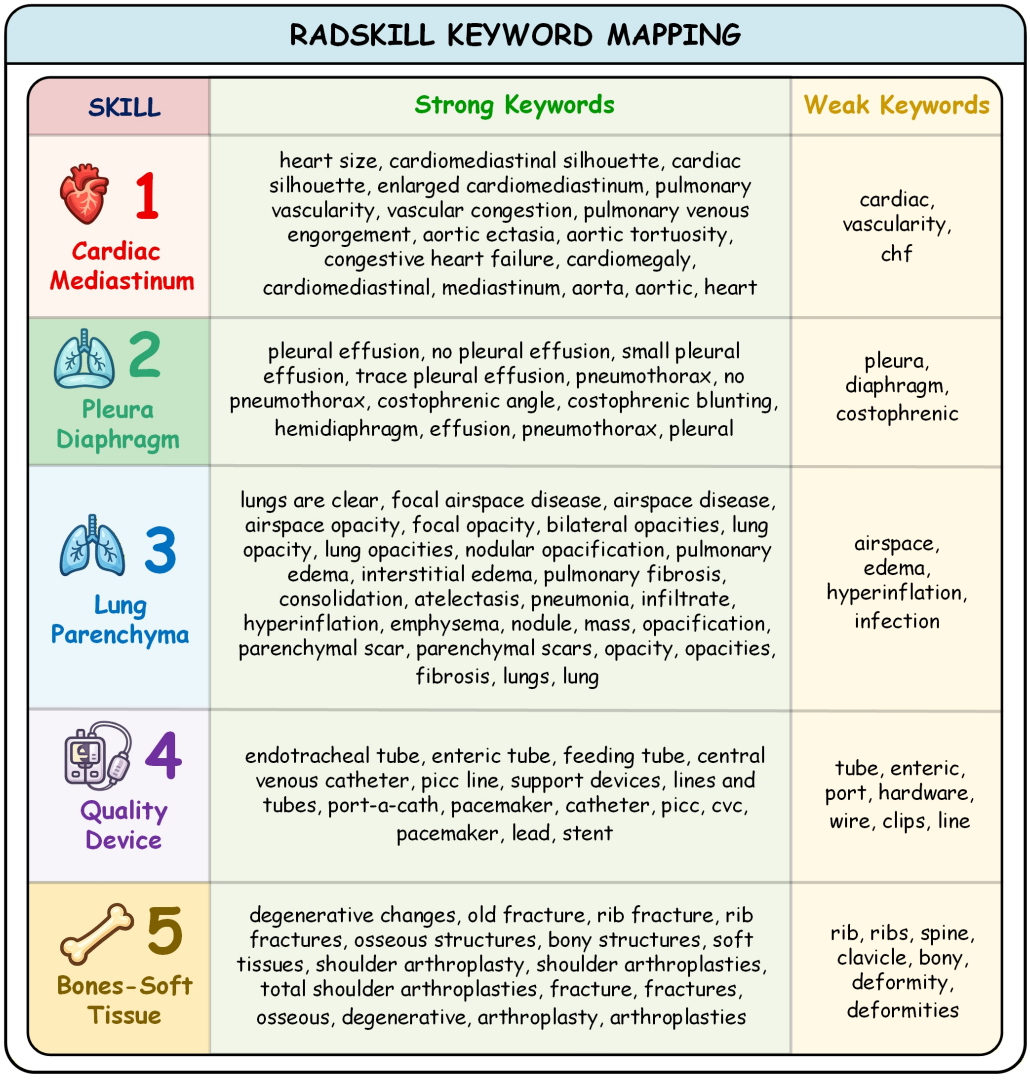


Figure C.8: Two-level medical dictionary for RadSkill mapping.

C.3 PSY-PO Implementation Details

Hallucination Candidate Construction and 3PL Fitting. For each ground-truth report, we construct a hallucination candidate pool containing 10 candidate hallucination reports independently sampled from Qwen3.5-9B using the Hallucination Injection Prompt shown in Figure C.6. The decoding settings are temperature $= 0.7$, maximum generated tokens $= 512$, and presence penalty $= 1.0$. To ensure stable 3PL fitting, we estimate the discrimination, difficulty, and guessing parameters (a, b, and c) by maximizing the Bernoulli log-likelihood using L-BFGS-B [51], which employs line search rather than a fixed learning rate. Specifically, we initialize $a = 1$, $b = \mathrm{median}(\eta_q)$, and $c = 0.05$, with parameter bounds of $[0.01, 5]$, $[-4, 4]$, and $[0, 0.35]$, respectively. The estimated parameters are subsequently used to compute Fisher information for selecting informative hard negatives in PSY-PO.

C.4 Further Ablation Results

Performance of Skill-evolution Epochs. We evaluate the impact of the dynamic evolution loop in SECR with different skill-evolution epochs ($t = 0, 1, 3, 5$) on final report generation performance on IU X-Ray. The results in Table C.1 show that both natural language generation and clinical efficacy metrics steadily improve as t increases. With the dynamic evolution loop, the RRG baseline

continuously revises the description and workflow of the RadSkill repository by identifying failed skill invocation samples and regenerating the corresponding clinical reasoning trajectories, thereby providing higher-quality supervision signals. When $t = 5$, all metrics achieve the best performance, with BL-4 and F1 improving to 0.203 and 0.517, respectively. These results demonstrate that SECR can continuously refine skill representations and clinical reasoning workflows through failure feedback, which is effectively internalized by the RRG baseline into more robust diagnostic capability during subsequent report generation.

Table C.1: Performance of increasing skill-evolution epochs t on IU X-Ray.

Epoch t	BL-4	MTR	RG-L	P	R	F1
$t = 0$	0.195	0.209	0.361	0.531	0.476	0.501
$t = 1$	0.195	0.215	0.364	0.530	0.485	0.506
$t = 3$	0.200	0.216	0.368	0.540	0.489	0.513
$t = 5$ **(Ours)**	**0.203**	**0.227**	**0.374**	**0.548**	**0.489**	**0.517**

Effect of Different Checkpoint Numbers. Table C.2 reports the impact of the number of historical model checkpoints on PSY-PO performance based on the RRG baseline without SECR. In our implementation, the 15-epoch report-generation training saves one checkpoint every half epoch, yielding up to $N = 30$ historical checkpoints. The results show that both NLG and CE metrics steadily improve as N increases from 10 to 30, except for a slight ROUGE-L dip at $N = 20$. When N is small, the checkpoint group covers a limited ability range, making the 3PL model less stable in estimating the clinical discrimination $a_{z,k}$, clinical hallucination concealment $b_{z,k}$, and incidental resistance rate $c_{z,k}$ of candidate hallucination reports, which further affects the computation of Fisher information. Increasing the number of checkpoints provides a more complete response distribution, improves the stability of parameter estimation, and enables more accurate identification of the clinical blind spots of the current model. When $N = 30$, the model achieves the best performance across all metrics, indicating that sufficient historical checkpoints help reliably quantify the clinical misleadingness of candidate hallucination samples, thereby selecting more effective hard negative reports and enhancing preference optimization.

Table C.2: Impact of the number of historical model checkpoints N in PSY-PO on IU X-Ray.

Number of Checkpoints	BL-4	MTR	RG-L	P	R	F1
$N = 10$	0.209	0.211	0.369	0.521	0.449	0.482
$N = 20$	0.210	0.214	0.364	0.535	0.479	0.506
$N = 30$ **(Ours)**	**0.213**	**0.220**	**0.375**	**0.557**	**0.517**	**0.536**

Performance of Cross-domain Generalization. Table C.3 evaluates the cross-domain generalization ability of the self-evolving clinical reasoning supervision constructed by SECR. In the cross-domain setting with MIMIC-CXR as the source domain and IU X-Ray as the target domain, "Baseline w/ SECR" outperforms Baseline on both NLG and CE metrics. The Baseline directly learns the mapping from images to the findings section, which is more susceptible to source-domain reporting styles and distribution shifts. In contrast, SECR first learns clinical reasoning optimized through dynamic evolution and then learns to generate the findings section, providing the model with more generalizable clinical reasoning supervision. Moreover, the model with SECR maintains consistent advantages in the target domain, suggesting that SECR not only improves source-domain performance but also captures transferable clinical reasoning patterns across datasets. These results demonstrate that SECR provides effective support for cross-domain radiology report generation.

Table C.3: Performance of cross-domain generalization from MIMIC-CXR to IU X-Ray. "Baseline w/ SECR" refers to the RRG baseline with SECR.

Models	BL-4	MTR	RG-L	P	R	F1
Baseline	0.072	0.146	0.247	0.510	0.528	0.519
Baseline w/ SECR	**0.086**	**0.148**	**0.267**	**0.536**	**0.530**	**0.533**

Performance of Different RRG Methods. To ensure a more controlled and fair comparison, we further reproduce several representative RRG methods on IU X-Ray using the same data preprocessing, Findings-only setting, decoding strategy, and evaluation scripts as our method. As shown in Table C.4, our method achieves the best performance across all three NLG metrics, with BLEU-4, METEOR, and ROUGE-L scores of 0.241, 0.230, and 0.395, respectively. Specifically, compared with the strongest reproduced baseline, COMG, our method achieves absolute improvements of 0.036, 0.012, and 0.011 on these three metrics, respectively, demonstrating consistent gains in report generation quality under the unified evaluation setting.

Table C.4: Comparison with representative RRG methods on IU X-Ray under a unified evaluation protocol.

Methods	BL-4	MTR	RG-L
CoMT	0.140	0.171	0.324
COMG	0.205	0.218	0.384
R2Gen-Mamba	0.168	0.201	0.376
CheXOne [49]	0.171	0.203	0.382
Ours	**0.241**	**0.230**	**0.395**

C.5 Further Qualitative Results

Visualization of SECR Dynamic Evolution. Figure C.9 shows the ability of SECR to correct failed skill invocation trajectories during dynamic evolution. As the evolution rounds proceed, SECR continuously refines the RadSkill repository by filtering failed samples, constructing failure sample pools, and extracting failure cues, thereby progressively improving skill invocation quality. Overall, the number of failed samples decreases from 1,272 to 68, the cumulative resolved ratio reaches 94.7%, and Skill F1 improves from 65.7% to 73.1%. These results indicate that SECR can continuously optimize skill representations and reasoning workflows through failure feedback, substantially reducing skill invocation mismatches and providing more reliable clinical reasoning supervision signals for subsequent SFT.

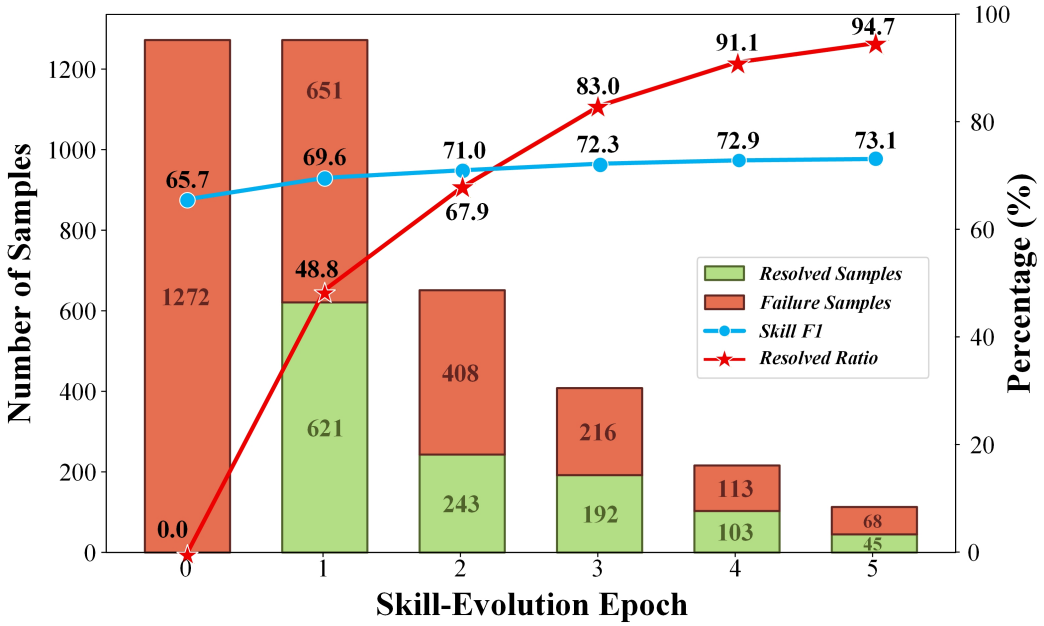


Figure C.9: Quantitative analysis of failure correction and skill invocation during SECR evolution. "Skill F1" denotes the F1 score of skill invocation correctness for samples in the current skill-evolution epoch. "Resolved Ratio" denotes the proportion of failed samples successfully corrected after evolution.

Visualization of RadSkill Evolution. To qualitatively analyze how RadSkills are refined in the dynamic evolution loop, Figure C.10 presents an evolution example of the Pleural Diaphragm skill from the initial version to the final evolved version. The initial version mainly focuses on basic abnormalities in the pleural cavity and diaphragmatic regions, and its diagnostic workflow relies primarily on common cues such as the apical pleural line. As evolution proceeds, missing information exposed by failed skill invocation trajectories is continuously fed back to the teacher model, enabling progressive refinement of the skill description and diagnostic workflow. The final version preserves the ability to identify basic abnormalities while further incorporating procedures for distinguishing abnormal anatomical changes and related pathological confounders, thereby covering diagnostic content insufficiently addressed by the initial skill. This process shows that SECR can close the loop between failure cases and the RadSkill repository by feeding back missing features and high-frequency diagnostic cues, continuously revising skill descriptions and refining diagnostic workflows, and ultimately transforming failure experience into more reliable explicit clinical reasoning trajectories.

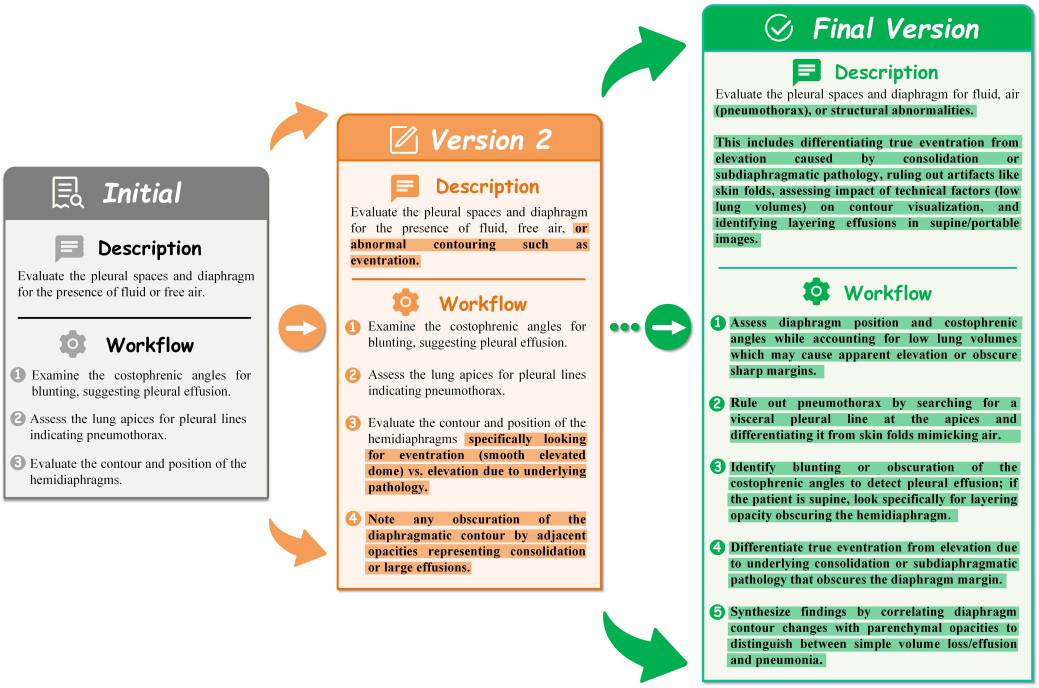


Figure C.10: Evolution of the Pleural Diaphragm RadSkill. Its description and workflow are progressively refined from the initial version to the final evolved version through the dynamic evolution loop.

Visualization of PSY-PO Candidate Selection. Figure C.11 visualizes the candidate negative report pool of PSY-PO under the 3PL model. Most candidate negatives are broadly distributed and exhibit relatively low discrimination and Fisher information, suggesting limited training utility for the current model. In contrast, the selected hard negative report is located near the current model ability $\eta_\theta = 0.55$ and shows high clinical discrimination and Fisher information. This indicates that the selected sample lies close to the model ability boundary and can more effectively expose clinically misleading hallucinations that are easily overlooked by the current model. These results demonstrate that PSY-PO does not rely on static sampling or random negative selection. Instead, it leverages collective responses from historical model checkpoints to dynamically identify more informative hard negative reports for subsequent preference optimization.

Qualitative Comparison of Hallucination Reports. Figure C.12 compares easy and hard hallucination reports with the ground-truth report on a MIMIC-CXR example. The case exhibits increased right mid-lung pneumonia, patchy opacity at the left base, a more prominent cardiac silhouette,

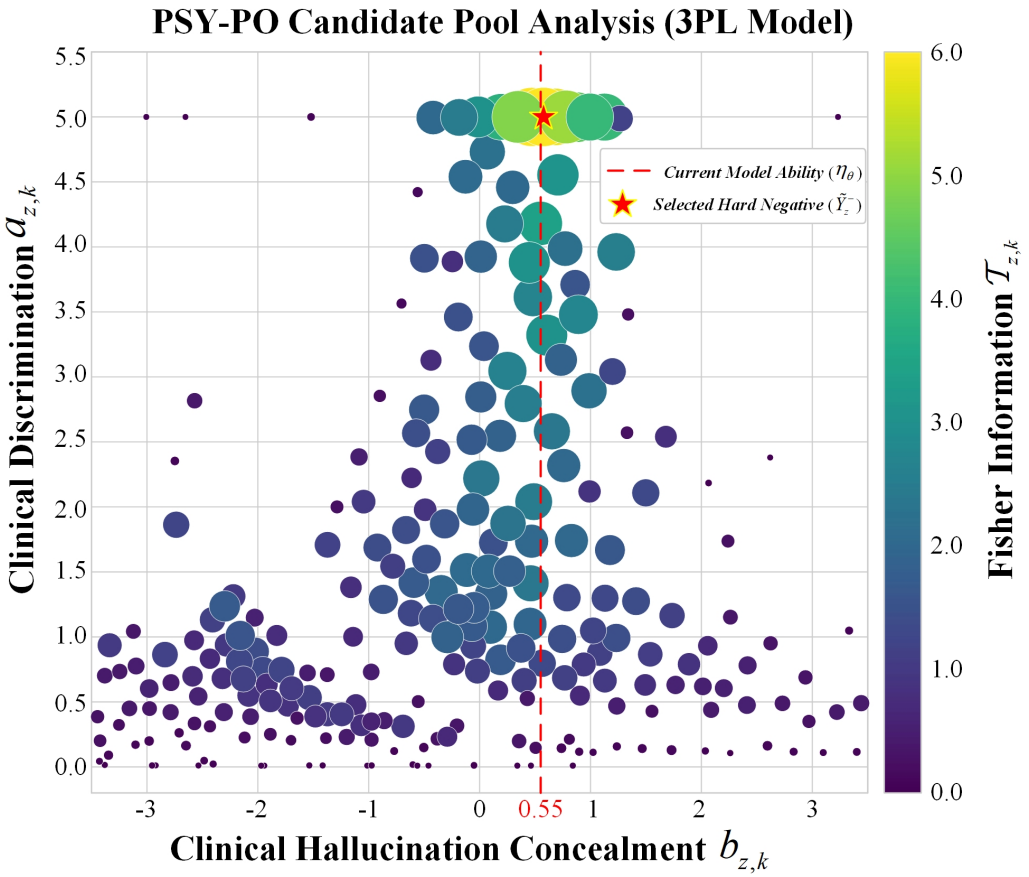


Figure C.11: Visualization of the PSY-PO candidate pool under the 3PL model on IU X-Ray. Each bubble denotes a candidate negative report, with its position determined by clinical hallucination concealment $b_{z,k}$ and clinical discrimination $a_{z,k}$. Bubble color and size encode the Fisher information $\mathcal{T}_{z,k}$. The red dashed line marks the current model ability η_θ, and the red star indicates the selected hard negative $\widetilde{Y}_z^-$, which lies near the model ability boundary and provides high Fisher information for preference optimization.

and bilateral small pleural effusions with compressive basilar atelectasis. The easy hallucination report introduces severe but visually unsupported abnormalities, such as a fluid-filled cavity, ruptured tension pneumatocele, malignant mass, and diaphragmatic rupture. These errors deviate substantially from the image evidence and can therefore be easily rejected, providing only coarse supervision. In contrast, the hard hallucination report remains highly consistent with the ground-truth report in sentence structure, disease category, and anatomical regions, while introducing subtle but clinically critical factual errors. For example, it replaces increased pneumonia with resolution, increased left-base opacity with decreased opacification, and a more prominent cardiac silhouette with a minimally prominent one. These errors are more challenging because the report appears fluent and clinically plausible, yet its key findings contradict the image evidence and the ground truth. This comparison suggests that conventionally sampled easy negatives provide limited supervision, whereas hard negatives selected by PSY-PO can offer more informative and discriminative preference signals for hallucination suppression.

Cross-Attention Visualization. Figure C.13 visualizes the cross-attention distributions of Baseline and Ours when generating descriptions related to right pleural effusion and compressive atelectasis. For right pleural effusion, the attention of Baseline is scattered and fails to focus sufficiently on the right pleural cavity and the adjacent costophrenic region, which leads to a false-negative statement in the generated report. In contrast, Ours places more attention on the right pleural region,

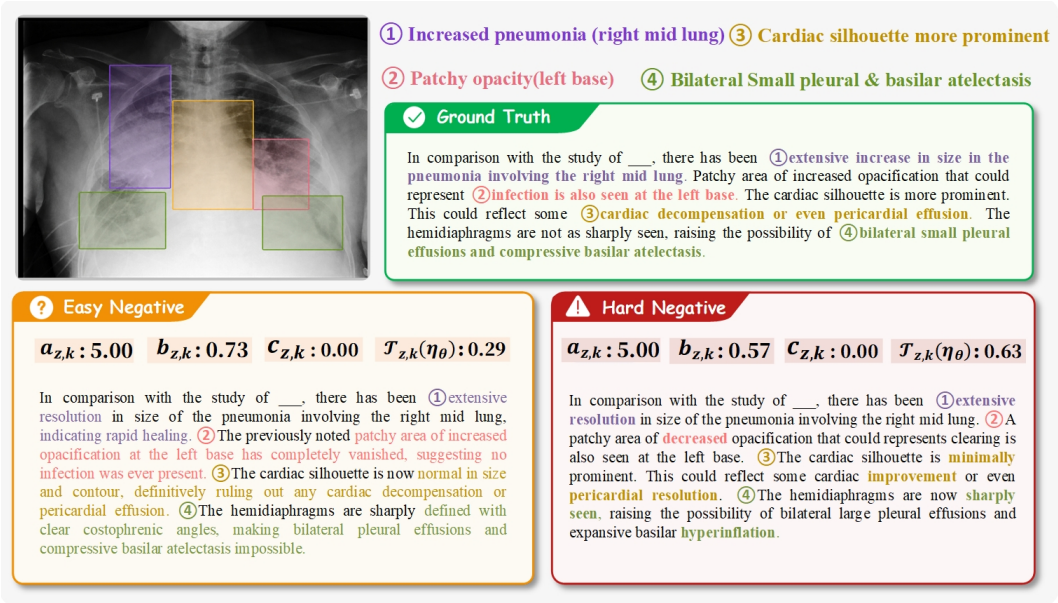


Figure C.12: Visualization of candidate hallucination samples under different sampling strategies on MIMIC-CXR. Green, orange, and red boxes denote the ground truth, easy negatives, and hard negatives, respectively.

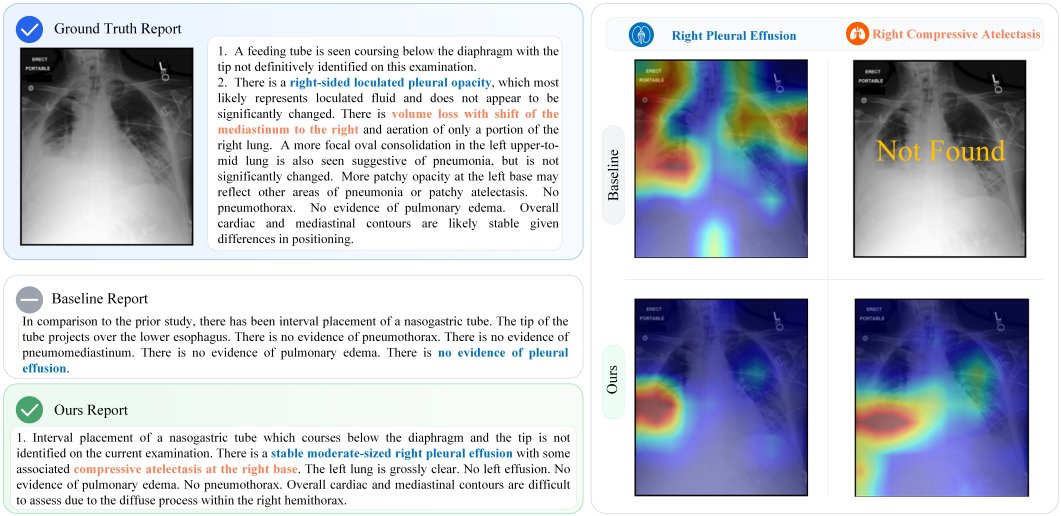


Figure C.13: Cross-attention visualization on MIMIC-CXR. The left panel highlights the main findings and their corresponding generated descriptions, while the right panels display the visual cross-attention distributions of both models when generating tokens associated with specific clinical findings.

making the reported positive finding more consistent with the visual evidence. For compressive atelectasis, Baseline fails to localize the abnormality effectively and therefore omits the corresponding description, whereas Ours attends more to the adjacent lung parenchyma, showing a more reasonable correspondence between image regions and textual findings. These results indicate that the explicit clinical reasoning supervision introduced by SECR, together with the preference constraints imposed by PSY-PO, can reduce false-negative generation and enhance cross-modal consistency between radiology reports and medical image evidence.

C.6 Robustness and Sensitivity Analysis

To evaluate the stability of our method, we repeat the main experiments using three different random seeds and report the mean and standard deviation of each metric.

Robustness of Ablation Modules and SFT Strategies. As shown in Tables C.5 and C.6, our method achieves the best mean performance across all NLG and CE metrics, with consistently small standard deviations. These results further validate the effectiveness and stability of the proposed modules and reasoning internalization strategy.

Table C.5: Robustness of module ablations on MIMIC-CXR across three random seeds. Standard deviations are shown in italics.

Different Modules	BL-4	MTR	RG-L	P	R	F1
Baseline	0.106 ± *0.003*	0.138 ± *0.002*	0.256 ± *0.003*	0.585 ± *0.005*	0.333 ± *0.003*	0.425 ± *0.003*
SECR w/o Evo	0.126 ± *0.003*	0.160 ± *0.002*	0.287 ± *0.003*	0.621 ± *0.003*	0.441 ± *0.004*	0.516 ± *0.003*
SECR	0.146 ± *0.003*	0.171 ± *0.004*	0.292 ± *0.001*	0.630 ± *0.004*	0.468 ± *0.007*	0.537 ± *0.003*
PSY-PO	0.155 ± *0.003*	0.175 ± *0.003*	0.303 ± *0.004*	0.617 ± *0.003*	0.446 ± *0.003*	0.518 ± *0.003*
Ours	**0.165** ± ***0.003***	**0.181** ± ***0.002***	**0.310** ± ***0.004***	**0.636** ± ***0.008***	**0.476** ± ***0.010***	**0.544** ± ***0.007***

Table C.6: Comparison of SFT strategies on IU X-Ray across three random seeds. Standard deviations are shown in italics.

Strategy	BL-4	MTR	RG-L	P	R	F1
Rep SFT	0.173 ± *0.010*	0.203 ± *0.002*	0.348 ± *0.007*	0.490 ± *0.005*	0.426 ± *0.003*	0.456 ± *0.003*
Rea+Rep SFT	0.103 ± *0.005*	0.143 ± *0.007*	0.263 ± *0.009*	0.492 ± *0.008*	0.480 ± *0.002*	0.486 ± *0.004*
Rep → Rea SFT	0.118 ± *0.006*	0.200 ± *0.003*	0.314 ± *0.007*	0.536 ± *0.008*	0.376 ± *0.009*	0.442 ± *0.004*
Ours	**0.209** ± ***0.010***	**0.228** ± ***0.001***	**0.379** ± ***0.009***	**0.550** ± ***0.006***	**0.491** ± ***0.005***	**0.519** ± ***0.006***

Robustness of IRT Parameterizations. Table C.7 reports the repeated results of different IRT parameterizations. The 3PL model achieves the highest mean performance across all metrics with low variability, further confirming the effectiveness and stability of the proposed psychometric preference optimization.

Table C.7: Comparison of IRT parameterizations across three random seeds. Standard deviations are shown in italics.

IRT Parameterization	BL-4	MTR	RG-L	P	R	F1
1PL	0.204 ± *0.003*	0.214 ± *0.003*	0.369 ± *0.004*	0.542 ± *0.007*	0.500 ± *0.005*	0.520 ± *0.006*
2PL	0.208 ± *0.003*	0.217 ± *0.002*	0.373 ± *0.004*	0.548 ± *0.006*	0.510 ± *0.006*	0.528 ± *0.006*
3PL	**0.213** ± ***0.002***	**0.220** ± ***0.003***	**0.376** ± ***0.003***	**0.557** ± ***0.006***	**0.517** ± ***0.007***	**0.536** ± ***0.005***

Effect of Teacher Model Scale and Prompt Style. To evaluate the sensitivity of our method to teacher-model capacity and prompt design, we conduct experiments using different teacher-model scales and prompt styles. As shown in Table C.8, the 9B teacher model achieves better overall performance than the 4B model across NLG and CE metrics. Specifically, the 9B model achieves clinical F1 scores of 0.576 and 0.573 under the original and concise prompts, respectively, compared with 0.544 for the 4B model under both configurations. These results suggest that teacher-model capacity influences clinical reasoning generation and downstream RRG performance. In contrast, varying the prompt style under a fixed teacher model leads to only minor fluctuations, mainly reflecting slight trade-offs across NLG and CE metrics. Overall, our method maintains stable performance across the tested prompt configurations.

Table C.8: Sensitivity analysis of teacher-model scale and prompt style on IU X-Ray. "Concise" denotes a simplified prompt that preserves the task objective and output format while removing detailed generation rules, constraints, and in-context examples, retaining only the core task instructions.

Teacher Model	Prompt Style	BL-4	MTR	RG-L	P	R	F1
9B	Ours	0.241	0.230	0.395	0.603	0.551	0.576
9B	Concise	0.240	0.223	0.396	0.596	0.553	0.573
4B	Ours	0.233	0.219	0.386	0.561	0.528	0.544
4B	Concise	0.238	0.221	0.388	0.572	0.519	0.544

C.7 Clinical Reliability Evaluation

Beyond conventional automatic evaluation metrics, assessing the clinical reliability of radiology report generation requires examining both the clinical validity of generated reports and the reliability of SECR reasoning trajectories. Accordingly, we conduct complementary evaluations at the report and reasoning levels to assess the clinical reliability of our method.

Radiologist Evaluation Setup. We conducted expert evaluations involving five radiologists on MIMIC-CXR and IU X-Ray. For generated-report evaluation, we randomly sampled 50 cases from each dataset based on CheXbert annotations, covering normal and abnormal cases as well as common and rare findings. These annotations were used solely for case selection and were not involved in model training or hyperparameter tuning. Radiologists independently assessed hallucination severity and clinical usefulness with access to the corresponding chest X-rays. Hallucination severity was graded as Correct Statement = 0, Mild Hallucination = 0.4, Moderate Hallucination = 0.7, and Severe Hallucination = 1. Clinical usefulness was graded as Clinically Acceptable = 1, Nearly Acceptable = 0.75, Moderately Acceptable = 0.5, Unacceptable = 0.25, and Unusable = 0. The methods and their outputs were anonymized and presented in randomized order, with radiologists blinded to the generating methods. For reasoning-trajectory evaluation, we randomly sampled 100 trajectories from the two datasets to assess reasoning quality and report-faithfulness before and after SECR, with all model configurations and experimental settings held fixed. Radiologists independently assessed the reasoning trajectories and corresponding reference reports without access to chest X-rays. Reasoning quality was graded as Sound = 1.0, Mostly Sound = 0.75, Partially Sound = 0.5, Flawed = 0.25, and Invalid Reasoning = 0. Report-faithfulness was graded as Fully Faithful = 1.0, Mostly Faithful = 0.6, Partially Faithful = 0.3, and Unfaithful = 0. Report-faithfulness measures consistency with the reference report rather than direct image-level faithfulness.

Rating Aggregation and Inter-Reader Agreement. Five radiologists independently evaluated the same sampled instances. For each method or ablation variant m, evaluation dimension q, and sampled instance i, we first averaged the ratings across the five radiologists, as shown in Eq. A.19:

$$\bar{s}_i^{(m,q)} = \frac{1}{5} \sum_{r=1}^{5} s_{i,r}^{(m,q)}. \tag{A.19}$$

where $s_{i,r}^{(m,q)}$ denotes the score assigned by radiologist r to instance i. The final evaluation score was then obtained by averaging the instance-level scores, as shown in Eq. A.20:

$$s^{(m,q)} = \frac{1}{N} \sum_{i=1}^{N} \bar{s}_i^{(m,q)}. \tag{A.20}$$

where $N = 50$ for each dataset in the generated-report evaluation and $N = 100$ for the reasoning-trajectory evaluation. To assess inter-reader agreement, we compute ordinal Krippendorff's $\alpha^{(m,q)}$ [16] separately for each method or ablation variant and evaluation dimension. Values closer to 1 indicate stronger agreement among radiologists, whereas $\alpha^{(m,q)} = 0$ indicates chance-level agreement.

Radiologist Assessments of Generated Reports. To address the limitations of relying solely on CheXbert label comparisons in assessing the quality of generated reports, we employed GREEN [30] to assess factual correctness alongside radiologist assessments of hallucination severity and clinical usefulness. As shown in Table C.9, our method achieves the highest GREEN scores of 0.371 and 0.712 and the highest clinical usefulness scores of 0.755 and 0.825 on MIMIC-CXR and IU X-Ray, respectively. Meanwhile, our method obtains the lowest hallucination severity scores of 0.295 on MIMIC-CXR and 0.120 on IU X-Ray. Compared with the baseline, our method consistently improves all three evaluation metrics on both datasets. These results demonstrate improved factual correctness, reduced hallucination severity, and enhanced clinical usefulness of the generated reports.

Table C.9: Clinical reliability evaluation of generated reports on MIMIC-CXR and IU X-Ray. GREEN scores and radiologist assessments of hallucination severity (↓) and clinical usefulness (↑) are reported for 50 sampled test cases from each dataset. Lower hallucination severity and higher GREEN scores and clinical usefulness indicate better performance.

Modules and Methods	MIMIC-CXR			IU X-Ray		
	GREEN ↑	Hallucination ↓	Usefulness ↑	GREEN ↑	Hallucination ↓	Usefulness ↑
Baseline	0.283	0.579	0.535	0.612	0.501	0.695
+SECR w/o Evo	0.313	0.546	0.590	0.630	0.427	0.650
+SECR	0.340	0.438	0.645	0.654	0.295	0.705
+PSY-PO	0.351	0.367	0.695	0.691	0.224	0.755
COMG+RL	0.298	0.331	0.620	0.664	0.188	0.710
R2Gen-Mamba	0.334	0.359	0.675	0.685	0.260	0.685
Ours	**0.371**	**0.295**	**0.755**	**0.712**	**0.120**	**0.825**

Inter-Reader Agreement on Report Assessments. To assess agreement among the five radiologists, we computed ordinal Krippendorff's α separately for hallucination severity and clinical usefulness under each compared method or ablation variant. A value closer to 1 indicates stronger inter-reader agreement, whereas $\alpha = 0$ indicates agreement equivalent to that expected under random ratings. As shown in Table C.10, agreement scores range from 0.655 to 0.742 for hallucination severity and from 0.660 to 0.741 for clinical usefulness. For our method, the agreement scores for hallucination severity and clinical usefulness are 0.655 and 0.662 on MIMIC-CXR, respectively, compared with 0.698 and 0.729 on IU X-Ray. Overall, these results indicate moderate inter-reader agreement across the two clinical assessment dimensions, providing additional evidence of consistency in the radiologist assessments.

Table C.10: Inter-reader agreement on radiologist assessments of generated reports. Ordinal Krippendorff's α for hallucination severity and clinical usefulness across five radiologists on MIMIC-CXR and IU X-Ray. Agreement was calculated separately for each method and evaluation dimension.

Methods	MIMIC-CXR		IU X-Ray	
	Hallucination Severity	Clinical Usefulness	Hallucination Severity	Clinical Usefulness
Baseline	0.716	0.741	0.672	0.731
+SECR **w/o** Evo	0.664	0.670	0.723	0.660
+SECR	0.742	0.701	0.681	0.670
+PSY-PO	0.703	0.681	0.708	0.718
COMG+RL	0.676	0.730	0.659	0.670
R2Gen-Mamba	0.729	0.671	0.733	0.698
Ours	**0.655**	**0.662**	**0.698**	**0.729**

Reasoning Quality and Report-Faithfulness Evaluation. To validate the reliability of SECR reasoning trajectories, we conducted automatic and manual evaluations before and after SECR. For automatic evaluation, GPT-5.5 was employed as an LLM judge to assess reasoning quality and report-faithfulness based on the reference radiology reports. For manual evaluation, radiologists independently assessed the same two dimensions. Reasoning quality evaluates the soundness of the reasoning process and whether its conclusions are adequately supported by relevant evidence, while report-faithfulness measures the consistency between reasoning trajectories and reference reports. As shown in Table C.11, SECR improves automatic reasoning quality from 7.552 to 8.829 and report-faithfulness from 6.676 to 8.270. The manual assessments confirm consistent improvements,

with reasoning quality increasing from 0.625 to 0.823 and report-faithfulness from 0.526 to 0.740. These results demonstrate that SECR improves reasoning soundness and alignment with reference reports, reducing the risk of internalizing inaccurate reasoning trajectories.

Table C.11: Reasoning quality and report-faithfulness before and after SECR. Automatic evaluation is conducted using GPT-5.5 on 500 cases with a rating scale ranging from 0 to 10, while manual evaluation was performed by five radiologists on 100 cases using a rating scale ranging from 0 to 1. Higher scores indicate better performance.

Reasoning Stage	**Automatic Analysis**		**Manual Audit**	
	Reasoning Quality	**Report Faithfulness**	**Reasoning Quality**	**Report Faithfulness**
Reasoning Chain **w/o** Evo	7.552	6.676	0.625	0.526
Reasoning Chain	8.829	8.270	0.823	0.740

Inter-Reader Agreement on Reasoning Assessments. To examine the consistency of manual reasoning assessments, we calculated ordinal Krippendorff's α separately for reasoning quality and report-faithfulness before and after SECR evolution. As shown in Table C.12, the agreement scores range from 0.712 to 0.739 for reasoning quality and from 0.733 to 0.748 for report-faithfulness. These results indicate moderate inter-reader agreement among the five radiologists across both evaluation dimensions, providing additional support for the reliability of the manual audit.

Table C.12: Inter-reader agreement on reasoning quality and report-faithfulness before and after SECR evolution. Ordinal Krippendorff's α across five radiologists evaluating 100 reasoning trajectories using the corresponding reference reports without access to chest X-rays.

Reasoning Stage	**Reasoning Quality Agreement**	**Report Faithfulness Agreement**
Reasoning Chain **w/o** Evo	0.712	0.748
Reasoning Chain	0.739	0.733

D Case Study

D.1 Examples on Explicit Clinical Reasoning

✔ Correct Clinical Reasoning ✖ Wrong Clinical Reasoning

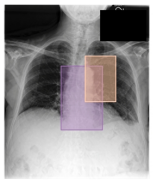
Input	Findings	Initial Clinical Reasoning	Final Clinical Reasoning
	Lung volumes remain low with suggestion of volume loss in both lower lobes most consistent with atelectasis. There is no evidence of pulmonary edema. Blunting of the left costophrenic angle suggests the presence of a small effusion or pleural thickening. Scattered calcified nodules in both lungs are unchanged since ___ and are therefore consistent with calcified granulomas. Overall cardiac and mediastinal contours are stable. Calcification of the aorta consistent with atherosclerosis.	First, the cardiac silhouette and mediastinal contours appear stable with no evidence of cardiomegaly or widening.✔ Second, parenchymal assessment reveals low lung volumes consistent with volume loss.✔ there are scattered calcified granulomas, which are unchanged and benign.✖ Third, pleural evaluation shows blunting of the left costophrenic angle, indicating a small effusion or pleural thickening on the left side.✔ No pulmonary edema is present,✔ and bony structures demonstrate atherosclerotic aortic calcification without acute fracture.✖ Conclusion: The findings point to bibasilar atelectasis with a concurrent small left pleural effusion/thickening in a clinical context of reduced inspiratory effort.✖	Analysis of lung volumes reveals low expansion, particularly in the lower lobes with volume loss consistent with bibasilar atelectasis.✔ The lung parenchyma otherwise appears clear; no focal consolidation or pulmonary edema is identified.✔ Pleural evaluation notes blunting of the left costophrenic angle, indicating a small left pleural effusion or pleural thickening, while the right angle is sharp.✔ The cardiac silhouette and mediastinal contours are stable without cardiomegaly.✔ Vascular calcification of the aorta suggests underlying atherosclerosis.✔ Scattered calcified nodules represent unchanged granulomas.✔ Conclusion: Findings indicate low lung volumes with bibasilar atelectasis, a probable small left pleural effusion, and stable chronic changes including aortic calcification and granulomas.✔

Figure D.1: Qualitative visualization of clinical reasoning evolution in SECR. The figure compares the initial and evolved clinical reasoning trajectories for the same chest X-ray sample. Ground-truth findings are color-coded to match the corresponding reasoning statements, while red crosses and green check marks indicate incorrect and correct reasoning steps, respectively.

Qualitative Comparison of Clinical Reasoning Evolution. Figure D.1 shows how the dynamic evolution loop in SECR improves the quality of clinical reasoning by refining an initial reasoning trajectory into a more complete and coherent evolved trajectory. The Initial CR identifies several local

findings, such as low lung volumes, blunting of the left costophrenic angle, scattered calcified nodules, and stable cardiac and mediastinal contours. However, its reasoning chain remains incomplete. It lacks stepwise verification of the lung parenchyma, pleural space, and chronic calcified changes, and loosely attributes bibasilar atelectasis and left pleural effusion or pleural thickening to reduced inspiratory effort, resulting in a gap between evidence and conclusion. In contrast, the Evolved CR first analyzes low lung volume and bibasilar atelectasis, then excludes focal consolidation and pulmonary edema, and further verifies the small left pleural effusion or pleural thickening with a sharp right costophrenic angle. It also supplements chronic findings such as calcified nodules and aortic atherosclerosis. This comparison demonstrates that SECR can use failed skill invocation trajectories to revise the descriptions and diagnostic procedures in RadSkill, transforming fragmented and weakly connected reasoning into a more evidence-complete and clinically coherent reasoning process. As a result, SECR provides more reliable explicit clinical reasoning supervision for subsequent report generation.

D.2 Examples on Final Radiology Report Generation

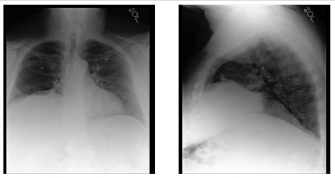
Ground-Truth	***Radiology Images***	***Baseline***
PA and lateral views of the chest provided. Midline sternotomy wires noted. Stable elevation of the right hemidiaphragm is again seen with chronic right basal atelectasis. Subtle retrocardiac linear density may represent focal areas of scarring as this appears unchanged from prior exam. No convincing signs of pneumonia or CHF. No large effusion or pneumothorax is seen. Cardiomediastinal silhouette is stable. Bony structures are intact. No free air below the right hemidiaphragm.	 *Frontal* *Lateral*	PA and lateral views of the chest provided. The patient is status post median sternotomy and CABG. The cardiomediastinal and hilar contours are normal. The lungs are clear. There is elevation of the right hemidiaphragm.
Baseline+SECR	***Baseline+PSY-PO***	***Ours***
PA and lateral views of the chest provided. Midline sternotomy wires again noted. There is persistent elevation of the right hemidiaphragm with right basal atelectasis. No focal consolidation concerning for pneumonia. No large effusion or pneumothorax. No signs of congestion or edema. Cardiomediastinal silhouette is stable. Bony structures are intact.	PA and lateral views of the chest provided. Midline sternotomy wires and mediastinal clips are again noted. There is stable elevation of the right hemidiaphragm. Lungs are clear. Cardiomediastinal silhouette is stable. No large effusion or pneumothorax. No convincing signs of pneumonia or edema. Bony structures are intact. No free air below the right hemidiaphragm.	PA and lateral views of the chest provided. Midline sternotomy wires and mediastinal clips are noted. There is persistent elevation of the right hemidiaphragm with right basal atelectasis. No convincing signs of pneumonia or edema. No large effusion or pneumothorax. Cardiomediastinal silhouette is stable. Bony structures are intact. No free air below the right hemidiaphragm.

Figure D.2: Qualitative comparison of generated reports on MIMIC-CXR. Color-coded sentences and phrases indicate matched clinical findings between the ground-truth report and reports generated by different models.

Qualitative Comparison of Generated Reports Across Ablation Modules. Figure D.2 presents a qualitative comparison between the ground-truth report and reports generated by different ablation variants on the MIMIC-CXR dataset. Colored spans highlight clinically relevant findings, including positive abnormalities, negative exclusions, and anatomical descriptions. The Baseline model captures only a few salient findings, such as normal cardiomediastinal and hilar contours, but misses several key descriptions, including midline sternotomy wires and a stable cardiomediastinal silhouette, indicating incomplete clinical coverage. When PSY-PO is introduced alone, Baseline+PSY-PO still incorrectly generates "the lungs are clear", leading to the omission of positive findings such as right basal atelectasis. This suggests that preference optimization alone may still lack sufficient explicit clinical reasoning support. Compared with Baseline+SECR, our full model provides more complete descriptions of implanted devices, such as mediastinal clips, and negative exclusions, such as no free air below the right hemidiaphragm. The generated report is also more consistent with the ground truth in both structure and semantics. These results show that our method better models positive abnormalities, negative exclusions, and anatomical status in chest radiographs, thereby generating more accurate, complete, and clinically reliable radiology reports.

E Limitations and Future Work

Although the proposed Look-Before-You-Leap framework achieves consistent performance improvements on two public chest X-ray datasets, several directions remain worth further exploration. First, the current work mainly focuses on chest X-ray report generation, with evaluation primarily conducted on the Findings section. In future work, we plan to extend this paradigm to more medical

imaging modalities, such as CT, MRI, and ultrasound, as well as multi-organ and multi-disease report generation tasks, to examine its generalization ability in more complex clinical settings.

Second, the SECR module relies on a teacher model to generate explicit clinical reasoning trajectories and continuously refines failed skill invocation trajectories through the dynamic evolution loop. Although this mechanism effectively improves the reliability of reasoning supervision signals, the medical knowledge coverage and generation stability of the teacher model may still affect the quality of the initial reasoning trajectories. Future work may introduce multi-teacher collaboration, domain-expert sampling review, or more fine-grained external medical knowledge constraints to further improve the accuracy and logical consistency of clinical reasoning trajectories. Additionally, the construction of the RadSkill supervision set relies on automatic medical entity and relation extraction using RadGraph, where subtle extraction errors may affect supervision reliability. Future work could incorporate multiple graph extractors, such as RadGraph and RadGraph-XL, for consistency checking, treating consistent results as high-confidence supervision while down-weighting or reviewing conflicting results. We have also supplemented our experiments with independent radiologist evaluations of generated reports and reasoning trajectories to further assess their clinical reliability.

Finally, the construction of preference data in PSY-PO inevitably involves some additional computational overhead, particularly during hallucination candidate generation and hard negative selection using historical model checkpoints. However, reusing checkpoints produced during GRG training reduces the need for independently trained examinee models, helping to moderate the additional cost. Future work could further improve the efficiency and robustness of preference data construction.

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 explicitly state our research scope and core contributions in the field of radiology report generation.

   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 the limitations of our proposed approach and potential future research directions in detail in Appendix E.

   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: Yes, the complete proofs for the main theoretical results are provided in Appendix A.

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 comprehensive information necessary for reproducing the main experimental results. The "Implementation Details" section specifies the model architecture, training paradigms, and evaluation metrics. Furthermore, Appendix C provides additional experimental details, including precise prompt templates.

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: [No]

   Justification: No. The source code is not open-sourced at this time. The experiments utilize completely public medical datasets (IU X-Ray and MIMIC-CXR), and the paper provides sufficiently detailed instructions on model training steps, hyperparameters, and evaluation protocols.

   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: We explicitly provide the data split strategies for the two public datasets in the "Datasets and Evaluation Metric" section, and describe the hyperparameter settings, optimizer type, and specific training details in the "Implementation Details" section.

   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: [Yes]

   Justification: We conduct repeated experiments using three different random seeds and report the mean and standard deviation of the evaluation metrics in Appendix C.6 (Tables C.7–C.9). The reported standard deviations reflect variability across random seeds under fixed experimental settings. Although formal statistical significance tests are not performed, these results provide an assessment of experimental variability and robustness.

   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: We detail the computational resources required for this research in the "Implementation Details" section.

   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: This research strictly adheres to the NeurIPS Code of Ethics. All medical imaging and report data used in our experiments are sourced from established, publicly available, and fully de-identified datasets, ensuring no privacy violations or conflicts of interest.

   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: We explore the positive societal impacts of automated radiology report generation in assisting clinical analysis.

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 research does not release models or data with a high risk of misuse, and thus does not involve such risks.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: All existing assets utilized in our study are properly cited in the References section. We have strictly complied with their respective credentialed access protocols, data usage agreements, and license requirements for research purposes.

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: [N/A]

    Justification: The research in this paper is based entirely on existing public datasets and does not introduce new datasets or other core new assets that require separate documentation.

    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: This study utilizes de-identified public datasets obtained through formal channels and does not involve crowdsourcing data collection or direct research on human subjects.

    Guidelines:

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

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

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

Answer: [N/A]

Justification: This study exclusively utilizes de-identified, publicly available datasets (MIMIC-CXR and IU X-Ray), and does not involve new human subjects research or crowdsourcing data collection. Therefore, no Institutional Review Board (IRB) approval is required.

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: Yes. We explicitly declare our usage of Large Language Models within the core methodology. Specifically, as detailed in the "Implementation Details" section, we utilize the pre-trained Qwen2.5-0.5B-Instruct model as the language decoder component of our Radiology Report Generation (RRG) baseline architecture, integrating it with the Rad-DINO vision encoder.

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.