FROM EVIDENCE ORCHESTRATION TO REASONING INTERNALIZATION FOR RADIOLOGY REPORT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Radiology report generation (RRG) methods often produce diagnostic findings without explicit clinical evidence, which may lead to factual hallucinations and lesion omissions. Although recent MLLM-based medical agents have begun to acquire clinical evidence through external tools, integrating conflicting clinical observations returned by different tools and mitigating error accumulation caused by early tool decisions remain challenging. Thus, we propose an evidence-informed RRG framework. Specifically, deliberative multi-tool evidence orchestration is first introduced, which employs Monte Carlo Tree Search (MCTS) to explore alternative tool-use trajectories and integrate outputs from different tools into structured clinical evidence. During MCTS decision-making, Counterfactual Search-Control Skill Evolution (CSC-SkillEvo) is further incorporated, which constructs counterfactual supervision signals from tool-use trajectories to refine tool-prioritization and search-termination decisions, thereby enabling holistic clinical evidence orchestration. Finally, we devise Curriculum-Annealed On-Policy Self-Distillation (CA-OPSD), which progressively restricts the clinical evidence accessible to the teacher to construct curriculum-based distillation signals, thereby enabling the RRG model to gradually internalize evidence-supported reasoning capabilities. Experiments on public datasets demonstrate that our framework consistently improves the clinical consistency of generated reports, thereby enhancing the reliability of RRG.

1 INTRODUCTION

Radiology Report Generation (RRG) (Chen et al., 2026; Liu et al., 2026c; Zhou et al., 2025) aims to automatically generate diagnostic reports from visual findings in radiological images. As RRG incorporates external medical knowledge, acquiring image-specific clinical evidence has become increasingly important. Recent MLLM-based medical agents (Feng et al., 2026; Ghezloo et al., 2025) dynamically invoke tools to acquire multi-source clinical evidence for medical image understanding and clinical decision-making. However, applying such tool-augmented paradigms to RRG remains challenging. Observations obtained by different tools around lesion regions may be inconsistent or even conflicting, increasing the difficulty of coordinating candidate tool decisions. Moreover, clinical evidence acquired through external tools is difficult to transform into stable report-generation capabilities of the RRG model. Therefore, reliable multi-tool evidence orchestration and effective reasoning internalization remain two key challenges for tool-augmented RRG.

Existing tool-augmented methods have evolved from sequential tool use based on interaction and feedback, such as ReAct (Yao et al., 2023b), to stronger planning paradigms, including explicit planning and subtask decomposition in Plan-and-Solve (Wang et al., 2023a), multi-step tool planning in ToolLLM (Qin et al., 2024), and tree-search-based decision-making over candidate action trajectories in Tree of Thoughts (ToT) (Yao et al., 2023a) and Language Agent Tree Search (LATS) (Zhou et al., 2024). Despite improved tool planning, observations obtained during tool use are still primarily used for reasoning and decision-making within the current task, rather than transformed into supervision signals for continually improving search-control capabilities. This limitation is particularly critical for RRG, where inconsistent observations from different tools for the same lesion and early decision errors can propagate through subsequent tool calls and degrade the acquired

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

clinical evidence. To address this issue, we propose **deliberative multi-tool evidence orchestration**, which leverages multi-branch outcomes from Monte Carlo Tree Search (MCTS) (Coulom, 2007) to construct counterfactual supervision signals for continually optimizing two agent skills: **tool-prioritization** and **search-termination**. The evolved skills are then fed back to guide subsequent tree search. Unlike approaches that make immediate tool decisions primarily based on the current search state or LLM priors, we further distill historical search experience into reusable search-control capabilities, enabling the tool-search process to progressively learn, much like radiologists accumulate diagnostic experience across cases, **which examination should be selected next and when further examination is no longer necessary**. Consequently, tree search outcomes are no longer used only for the decision, but instead support the continual evolution of search-control skills that guide subsequent searches, enabling more global orchestration of multi-tool clinical evidence.

High-quality clinical evidence does not necessarily translate into stable reasoning and generation capabilities in RRG models. A straightforward solution is to adopt training-free methods (Rieff et al., 2025; Bao et al., 2025; Lu et al., 2025), which explicitly provide external clinical evidence during inference and use chain-of-thought prompting to guide step-by-step reasoning. However, such methods remain dependent on additional clinical evidence and may suffer from increased inference complexity and distribution shifts introduced by heterogeneous evidence. Alternatively, training-based methods learn clinical reasoning capabilities through parameter updates, including supervised fine-tuning (SFT) (Ossowski et al., 2026; Zhuang et al., 2026), Group Relative Policy Optimization (GRPO) (Shao et al., 2024), and On-Policy Distillation (OPD) (Agarwal et al., 2024; Bousselham et al., 2026). Supervised fine-tuning under teacher forcing may not cover model-generated states, while sparse outcome-level rewards provide limited fine-grained supervision. In contrast, OPD provides token-level distribution supervision on student-generated rollouts, offering more fine-grained learning signals. Building on this paradigm, On-Policy Self-Distillation (OPSD) (Zhao et al., 2026a;b) further assigns teacher and student roles to the same model under different information conditions. Nevertheless, directly applying OPSD to RRG remains challenging. When the teacher policy conditions on clinical evidence acquired through search while the student policy is conditioned only on the original image, a substantial conditioning gap arises, making it difficult to stably transfer evidence-supported clinical reasoning to the student. Therefore, **we propose Curriculum-Annealed On-Policy Self-Distillation (CA-OPSD), which introduces the easy-to-hard principle of curriculum learning into OPSD.** By progressively reducing the clinical evidence available to the teacher policy, CA-OPSD constructs a progressive curriculum schedule for distillation from strong to weak evidence support, enabling the RRG model to progressively absorb evidence-supported clinical reasoning along its own generated trajectories and ultimately internalize it into stable report generation capabilities that rely only on the original image.

In summary, the proposed evidence-informed RRG framework establishes a complete pipeline centered on the orchestration and internalization of clinical evidence. For evidence orchestration, the proposed approach does not simply expand clinical evidence through tool invocation; instead, it transforms search feedback generated during tool use into supervision signals that promote the self-evolution of search-control capabilities. To further transfer clinical evidence into report generation capabilities, the diversity of clinical evidence encountered during search is fully exploited to construct a progressive distillation curriculum, enabling the RRG model to gradually transition from explicit evidence utilization to the formation of internal reasoning capabilities. To the best of our knowledge, this is the first framework in RRG to unify multi-tool clinical evidence orchestration with curriculum-based OPSD. The main contributions are summarized as follows:

1. We propose deliberative multi-tool evidence orchestration, which exploits diverse MCTS branches to construct counterfactual supervision and continuously refine tool-prioritization and search-termination skills via CSC-SkillEvo.
2. We propose CA-OPSD, which treats clinical evidence as privileged information and progressively anneals its availability to internalize evidence-supported reasoning into report-generation capabilities that rely only on the original image.
3. Experiments on the public MIMIC-CXR (Johnson et al., 2019) and IU X-Ray (Demner-Fushman et al., 2016) datasets demonstrate that the proposed framework effectively improves the clinical consistency and factual reliability of generated reports.

2 Related Work

Radiology Report Generation aims to extract clinically relevant findings from radiological images and generate diagnostically accurate and semantically coherent reports. Early RRG studies mainly adopted encoder–decoder architectures, evolving from CNN-RNN models (Li et al., 2018; Jing et al., 2019) with attention mechanisms to Transformer-based designs (Chen et al., 2020; Zhu et al., 2023) for cross-modal information fusion between images and long-sequence reports. More recently, LLM-based decoders (Wang et al., 2025c;b; Che et al., 2026) have been introduced to project medical image features into the semantic space of LLMs through vision-language alignment, leveraging their linguistic knowledge and text-generation capabilities to improve report quality. Alongside the evolution of RRG architectures, some studies have incorporated external medical knowledge, including knowledge graphs (Liu et al., 2025b), similar report retrieval (Liu et al., 2026b;a), and retrieval-augmented generation (RAG) (Yoon et al., 2026), to provide additional medical evidence. Beyond predefined medical knowledge, recent medical agents (Feng et al., 2026; Ghezloo et al., 2025) have adopted tool-augmented paradigms that actively invoke specialized tools to acquire clinical evidence. However, existing tool-augmented methods still largely rely on local step-by-step decisions and lack global evaluation of multi-tool trajectories and their evidence combinations. This hinders the acquisition of reliable and complementary external medical evidence around lesion regions, compromising evidence completeness and reliability.

On-Policy Distillation performs token-level distribution supervision from a stronger teacher model on rollouts generated by the student policy itself (Agarwal et al., 2024; Bousselham et al., 2026; Dai et al., 2025), alleviating the distribution shift between training data and the actual inference states of the student in conventional off-policy distillation (Xu et al., 2026; Nicolicioiu et al., 2026). To reduce dependence on an independent teacher model, recent studies (Zhao et al., 2026a;b; Shrestha & Tessier, 2026) have explored OPSD, where the same model is instantiated as teacher and student under different conditioning contexts for self-distillation. Subsequent works improve the reliability of teacher token supervision (Liu et al., 2026d; Shen et al., 2026) and token-level credit assignment (Nguyen et al., 2026; Hou et al., 2026) to mitigate teacher noise. However, these studies mainly focus on general reasoning scenarios, while OPSD remains underexplored in medical artificial intelligence. A key challenge is that intermediate diagnostic evidence containing rich lesion and abnormality information, when used as privileged information for the teacher, can substantially enlarge the teacher-student conditioning gap. Since the student relies only on the original image, this gap makes it difficult to absorb evidence-supported reasoning capabilities from the teacher. Therefore, mitigating the teacher-student conditioning gap induced by external clinical evidence remains a vital problem for applying OPSD to medical artificial intelligence.

3 Methodology

3.1 Overview

Figure 1 illustrates the overall pipeline of the proposed evidence-informed RRG framework, which consists of two stages: deliberative multi-tool evidence orchestration and CA-OPSD. In the former stage, the CSC-SkillEvo, equipped with a tool-prioritization skill $\mathbb{S}_{tp}$ and a search-termination skill $\mathbb{S}_{st}$, is embedded into the MCTS process to iteratively optimize tool selection and search-termination. Given a radiological image x, an MLLM is employed as the decision agent $\mathcal{A}_\theta$, which uses MCTS to explore multiple tool-use trajectories over the tool space of the radiology tool library $\mathbb{T}_{\text{rad}}$. During tree search, the tool-use trajectories are used to construct two types of counterfactual contexts for tool-prioritization and search-termination, which are fed into a skill optimizer $\mathcal{A}_s$ to perform the CSC-SkillEvo loop. Meanwhile, the evolved $\mathbb{S}_{tp}$ and $\mathbb{S}_{st}$ are fed back into subsequent tree searches, enabling holistic clinical evidence orchestration. In the CA-OPSD stage, the positive evidence E_x^+ and the hard-negative evidence E_x^- are extracted from all trajectories of each search tree and combined with the corresponding ground-truth (GT) report to construct a progressive curriculum schedule $\mathcal{C}$ as privileged information. These curriculums correspond to evidence discrimination $\mathcal{C}_{ed}$, evidence grounding $\mathcal{C}_{eg}$, and reasoning consolidation $\mathcal{C}_{rc}$, respectively. For each curriculum $\mathcal{C}_k \in \mathcal{C}$, the corresponding teacher policy is instantiated, and its token-level distribution supervision progressively internalizes evidence-supported clinical reasoning capabilities into the student policy.

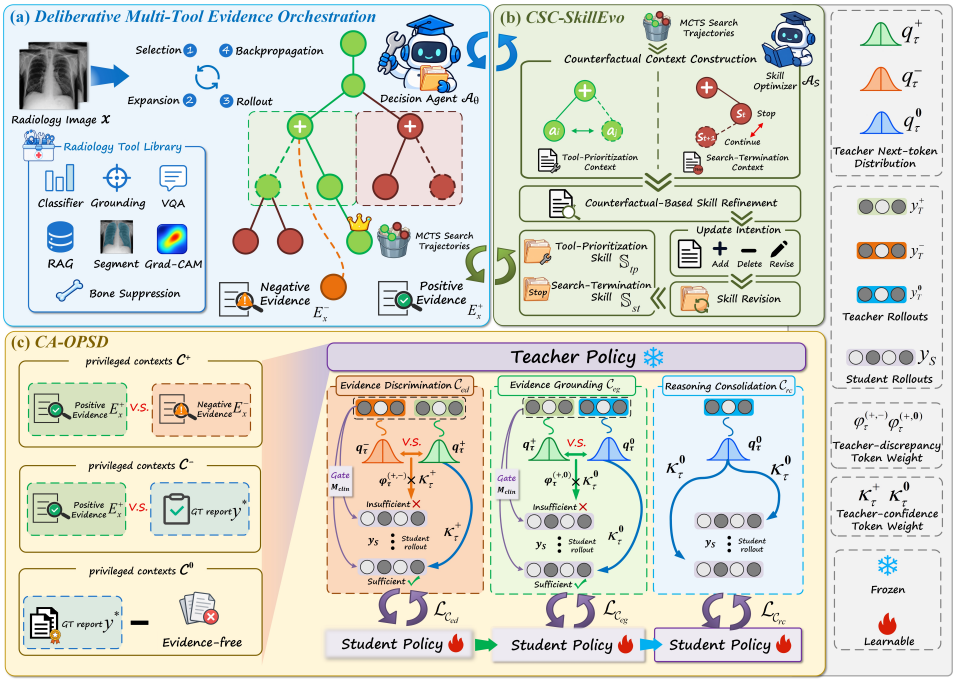

Figure 1: Pipeline of the proposed evidence-informed RRG framework.

3.2 Deliberative Multi-Tool Evidence Orchestration

To support clinical evidence acquisition in RRG, we construct a radiology tool library $\mathbb{T}_{\text{rad}}$ organized into three functional categories: global diagnostic tools (Classifier and RAG), localized analysis tools (Grounding, Segment, and Bone Suppression), and finding-level verification tools (VQA and Grad-CAM). We formulate clinical evidence acquisition as a tool-planning tree search problem, where the decision agent $\mathcal{A}_\theta$ selects candidate tools from the current state and MCTS explores and globally evaluates tool-use trajectories. Our approach organizes tree search into multiple search rounds and applies CSC-SkillEvo between rounds **(see Section 3.3)** to iteratively refine tool-prioritization and search-termination skills from historical search experience. Our evidence orchestration method, as illustrated in Figure 1(a), does not simply combine MCTS with CSC-SkillEvo. Instead, it forms a closed loop of mutual feedback between tree search and skill evolution. For each search round, we construct a multi-tool search tree whose nodes correspond to evidence states s_t and edges represent tool actions a_t. Each state contains prior tool-use actions and their observations. The Classifier tool initializes the root by identifying positive findings as anchor findings for subsequent tool selection and search-termination decisions. Evidence orchestration then iterates through four stages: Selection, Expansion, Rollout, and Backpropagation.

Selection. Each search round starts from the root node corresponding to the initial state s_0 and iteratively traverses the current search tree by selecting child nodes until reaching an expandable frontier node. For the state s_t corresponding to any non-leaf node, we adopt the PUCT criterion (Rosin, 2011) to select the most promising action a_t^* from the executable candidate action set Eq. 1:

$$a_t^* = \arg\max_{a \in \Upsilon(s_t)} \left[Q(s_t, a) + \tilde{c} \cdot P(a \mid s_t) \frac{\sqrt{N(s_t)}}{1 + N(s_t, a)} \right], \tag{1}$$

where $Q(s_t, a)$ denotes the mean action value of action a at state s_t. $N(s_t, a)$ and $N(s_t)$ denote the edge visit count and node visit count, respectively. $\tilde{c}$ is a constant that balances the exploitation of high-value branches with the exploration of promising but underexplored branches. $P(a \mid s_t)$ denotes the normalized result of the tool-priority score produced by $\mathcal{A}_\theta$ using the tool-prioritization skill $\mathbb{S}_{tp}$ for candidate action a at state s_t during Expansion.

Expansion. Upon reaching an expandable frontier node, the decision agent $\mathcal{A}_\theta$ generates the executable candidate action set $\Upsilon(s_t)$ based on the current state s_t. To prioritize branches with greater potential value under a limited tool-use budget, we use the current tool-prioritization skill $\mathbb{S}_{tp}$ to assign a tool-priority score to each candidate action. We then retain the $Top3$ candidate actions with the highest scores and expand their corresponding child nodes in the search tree.

Rollout. Among the $Top3$ candidates retained during Expansion, the decision agent $\mathcal{A}_\theta$ executes the unvisited action a_t with the highest $P(a \mid s_t)$. The resulting observation is incorporated into the search context and summarized by a *summary prompt* into structured clinical evidence E, updating the state to s_{t+1}. CheXbert extracts CheXpert-14 labels from E and the GT report, whose F1 score defines the state value $V(s_{t+1})$. Once each anchor finding has at least one valid observation, the search-termination skill $\mathbb{S}_{st}$ determines whether to terminate the search or continue exploration.

Backpropagation. After obtaining the state value $V(s_{t+1})$, we backpropagate it along the search path from the corresponding child node to the root. The edge visit count $N(s,a)$, cumulative value $W(s,a)$, and mean action value $Q(s,a)$ are updated accordingly. The updated $Q(s_t,a)$, together with $P(a \mid s_t)$, is then used in subsequent PUCT computations, allowing the clinical evidence value obtained from the current rollout to propagate back to earlier tool decisions and guiding subsequent search toward branches with greater potential clinical evidence value.

3.3 Counterfactual Search-Control Skill Evolution

To enable MCTS to acquire planning capabilities for tool selection and search termination over multiple search rounds, this section details the implementation of CSC-SkillEvo in Figure 1(b).

Counterfactual Context Construction. Given the predefined tool-prioritization skill $\mathbb{S}_{tp}$ and search-termination skill $\mathbb{S}_{st}$, their fixed rules cannot adapt tool priorities and search-termination decisions to actual search outcomes. We construct counterfactual data by comparing the realized search values of the current and alternative decisions to determine whether the current decision should be revised or retained, providing forward-looking counterfactual supervision for CSC-SkillEvo.

For tool-prioritization counterfactuals, we randomly sample a search node with at least two child nodes from the tree generated in the r-th MCTS round, with state s_t. The action a_i prioritized by $\mathbb{S}_{tp}$ is treated as the current decision, while a_j from another explored child branch serves as the alternative. We obtain their normalized tool-priority scores $P(a_i|s_t)$ and $P(a_j|s_t)$, together with the mean action values $Q(s_t,a_i)$ and $Q(s_t,a_j)$. The sample is labeled Revise if $P(a_i|s_t) \geq P(a_j|s_t)$ but $Q(s_t,a_i) < Q(s_t,a_j)$, indicating that the alternative decision achieves a higher search value despite receiving a lower priority. Otherwise, the sample is labeled Keep. This yields the round-r tool-prioritization counterfactual data $\mathcal{D}^r_{tp} = \{M(s_t); a_i; a_j; \psi\}$, where $M(s_t)$ denotes the search context and $\psi \in \{\text{Revise}, \text{Keep}\}$ is the update label. For search-termination counterfactuals, we randomly select a leaf node where the decision agent $\mathcal{A}_\theta$, guided by $\mathbb{S}_{st}$, terminates the search in the r-th MCTS round. Its state s_t is treated as the current termination decision. We then perform a one-step counterfactual expansion to obtain a child state s_{t+1}, treating continued search as the alternative decision. The sample is labeled Revise if $V(s_{t+1}) > V(s_t)$, indicating that continuing the search leads to a higher search value. Otherwise, it is labeled Keep. This yields the round-r search-termination counterfactual data $\mathcal{D}^r_{st} = \{M(s_t); s_t; s_{t+1}; \psi\}$.

Counterfactual-Based Skill Refinement. The two types of counterfactual contexts constructed above are further transformed into reusable agent skills. Unlike existing skill self-evolution methods that directly feed feedback contexts into an LLM for one-shot skill revision, our refinement process explicitly models error attribution, effective decision abstraction, and holistic rule refinement. Inspired by the iterative improvement process of professional decision-making through case review and experience abstraction, we design counterfactual-based skill refinement. A set of specialized prompts guides the skill optimizer $\mathcal{A}_s$ to identify error patterns and effective decision rationales from concrete counterfactual cases, formulate structured update intentions, and finally refine the current skill. In this way, counterfactual experiences are distilled into reusable tool-prioritization and search-termination rules, enabling the self-evolution of agent skills. For the tool-prioritization skill $\mathbb{S}^r_{tp}$ and search-termination skill $\mathbb{S}^r_{st}$ in the r-th search round, the corresponding counterfactual data $\mathcal{D}^r_{tp}$ and $\mathcal{D}^r_{st}$ are grouped into mini-batches according to their update labels. A *diagnosis prompt* guides $\mathcal{A}_s$ to contrastively analyze Revise-labeled and Keep-labeled samples. Revise-labeled samples are used to identify rule deficiencies that cause inappropriate tool-prioritization or search-

termination decisions, while Keep-labeled samples are used to extract effective decision rationales that should be preserved, producing diagnosis logs. Next, an *intention prompt* combines the diagnosis logs with the current skill to generate structured update intentions. These intentions specify three rule-level operations: add, delete, and revise, which respectively introduce missing rules, remove redundant or conflicting rules, and adjust the applicability conditions and scopes of existing rules. Finally, a *revision prompt* guides $\mathcal{A}_s$ to holistically refine the current skill according to the generated update intentions, producing the corresponding candidate skills $\tilde{\mathbb{S}}_{tp}^{r}$ and $\tilde{\mathbb{S}}_{st}^{r}$. This process realizes an iterative refinement paradigm that transforms counterfactual experience into holistic skill updates.

Skill Validation Gate. To ensure that counterfactual skill updates can effectively improve MCTS performance, we introduce a skill validation gate to screen the candidate skills $\tilde{\mathbb{S}}_{tp}^{r}$ and $\tilde{\mathbb{S}}_{st}^{r}$. For either $\tilde{\mathbb{S}}_{tp}^{r}$ or $\tilde{\mathbb{S}}_{st}^{r}$ in the r-th search round, we conduct MCTS with the current skill and the corresponding candidate skill on a predefined radiological image validation set under identical search configurations. The candidate skill is accepted only when the best search trajectories obtained with the candidate skill achieve a higher average final CheXpert F1 score than those obtained with the current skill. Otherwise, the current skill is retained. The validated skills $\mathbb{S}_{tp}^{(r+1)}$ and $\mathbb{S}_{st}^{(r+1)}$ are then used in the next MCTS search round, where new counterfactual contexts are constructed from the resulting search trees to continually drive CSC-SkillEvo.

After completing the three stages of CSC-SkillEvo, the evolved tool-prioritization and search-termination skills serve as reusable agent skill modules for unseen radiological images and guide MCTS to generate corresponding tool-use trajectories. Subsequently, guided by a *summary prompt*, the decision agent $\mathcal{A}_\theta$ integrates the tool observations and search context from a selected tool-use trajectory into textual clinical evidence **(see Appendix B.2 for prompt details of the three stages).**

3.4 Curriculum-Annealed On-Policy Self-Distillation for Reasoning Internalization

Benefiting from multi-tool evidence orchestration, we obtain external clinical evidence from selected tool-use trajectories. However, internalizing such evidence into the diagnostic reasoning capability of the RRG model while reducing reliance on it during inference remains challenging. Inspired by OPSD, the teacher policy and student policy are instantiated from the same GT-supervised RRG model π_θ. During distillation, theteacher policy is kept frozen, while only the student policy is updated. The teacher policy additionally conditions on external clinical evidence as privileged context and provides token-level distribution supervision on student rollouts. This increases the teacher-student policy gap and makes evidence-supported diagnostic reasoning difficult to absorb. Thus, as shown in Figure 1(c), we propose CA-OPSD, which adopts an annealed curriculum schedule to progressively remove clinical evidence from the teacher's privileged context. The three curriculum settings correspond to evidence discrimination $\mathcal{C}_{ed}$, evidence grounding $\mathcal{C}_{eg}$, and reasoning consolidation $\mathcal{C}_{rc}$, progressively transitioning training toward an evidence-free setting aligned with inference.

Curriculum Setup. Given an image x and its GT report y^*, the student policy generates an on-policy rollout $y_s \sim \pi_S(\cdot|x)$ with next-token distribution $p_\tau = \pi_S(\cdot|x, y_{s<\tau})$. We extract the clinical evidence from the best MCTS trajectory as positive evidence E_x^+. Additional trajectories with high textual similarity to the GT report but misleading clinical errors provide hard-negative evidence E_x^- **(see Appendix B.3).** We then construct three teacher privileged contexts: $c^+ = (y^*, E_x^+)$, $c^- = (y^*, E_x^-)$, and $c^0 = (y^*)$ π_T^+, negative teacher policy π_T^-, and GT report teacher policy π_T^0. All three teacher policies are evaluated on the same student rollout, with $q_\tau^\alpha = \pi_T^\alpha(\cdot|x, c^\alpha, y_{s<\tau})$, where $\alpha \in \{+, -, 0\}$.

Evidence Discrimination and Evidence Grounding. To enable the student policy to progressively internalize diagnostic information from clinical evidence, the first two curriculum stages of CA-OPSD focus on evidence discrimination and evidence grounding, respectively. In $\mathcal{C}_{ed}$, the student first learns to distinguish reliable from misleading diagnostic information by comparing E_x^+ and E_x^-. Subsequently, $\mathcal{C}_{eg}$ removes E_x^- and retains only E_x^+ as additional clinical evidence, further internalizing the incremental diagnostic information provided by E_x^+ relative to the GT report condition into the student policy. Both curriculum stages use the positive teacher policy π_T^+ as the final distillation target, while π_T^- and π_T^0 are used only as comparison signals in $\mathcal{C}_{ed}$ and $\mathcal{C}_{eg}$, respectively.

The teacher policy additionally receives intermediate clinical evidence, whereas the student policy is conditioned only on the original image. This creates a substantial teacher-student policy gap. Applying uniform strong supervision to all student rollouts makes it difficult to account for intra-curriculum differences in learning states and inter-curriculum differences in objectives. Therefore, before entering each curriculum stage, CA-OPSD first evaluates the tendency of each student rollout toward the current curriculum objective, and then adaptively adjusts the token-level supervision weights based on this tendency. If a student rollout exhibits the corresponding curriculum-objective tendency, token weighting is determined by the prediction reliability of the teacher policy at different generation positions. Specifically, we define the token-level confidence using normalized entropy as $\kappa_\tau^\alpha = 1 - H(q_\tau^\alpha)/\log|\mathcal{V}|$, where $H(\cdot)$ denotes entropy and $|\mathcal{V}|$ is the vocabulary size. In contrast, if a student rollout does not yet exhibit a sufficient tendency toward the curriculum objective, we further incorporate the token-level discrepancy between different teacher distributions to strengthen curriculum-specific supervision. Accordingly, we define the token-level teacher discrepancy as $\varphi_\tau^{(\alpha_1,\alpha_2)} = D_{\mathrm{JS}}(q_\tau^{\alpha_1} \parallel q_\tau^{\alpha_2})/\log 2$, which measures the prediction discrepancy between teacher distributions under different privileged contexts at the same generation position, where $\alpha_1, \alpha_2 \in \{+, -, 0\}$.

We model the tendency of each student rollout y_s toward the current curriculum objective through its report-level relative clinical consistency with different teacher rollouts. Let y_T^+, y_T^-, and y_T^0 denote the rollouts generated by π_T^+, π_T^-, and π_T^0, respectively. We measure the clinical consistency between two reports using $M_{\mathrm{clin}}(y^{\alpha_1}, y^{\alpha_2}) = [\text{BL-4}(y^{\alpha_1}, y^{\alpha_2}) + \text{F1}(y^{\alpha_1}, y^{\alpha_2})]/2$. For $\mathcal{C}_{ed}$, if a student rollout satisfies $M_{\mathrm{clin}}(y_s, y_T^+) \le M_{\mathrm{clin}}(y_s, y_T^-)$, this indicates that the rollout has not yet exhibited a stronger tendency toward the positive teacher rollout than toward the negative teacher rollout, and therefore has not sufficiently achieved the objective of evidence discrimination. We therefore use $\varphi_\tau^{(+,-)}$ to strengthen the token-level distillation signals at positions where the two evidence conditions induce different diagnostic predictions. For $\mathcal{C}_{eg}$, if $M_{\mathrm{clin}}(y_s, y_T^+) \le M_{\mathrm{clin}}(y_s, y_T^0)$, this indicates that the rollout has not yet exhibited a stronger tendency toward the positive teacher rollout than toward the GT-only teacher rollout, and therefore has not sufficiently absorbed the incremental diagnostic information provided by E_x^+. We therefore use $\varphi_\tau^{(+,0)}$ to strengthen the token-level distillation signals at the corresponding positions. Otherwise, for student rollouts that already exhibit the corresponding curriculum-objective tendency, token-level weighting relies only on the prediction confidence of the teacher policy. Let $z \in \{ed, eg\}$ denote the curriculum index. The resulting token weight is uniformly defined in Eq. 2:

$$w_\tau^z = \begin{cases} \kappa_\tau^+ \varphi_\tau^{(+,-)}, & z = ed,\ M_{\mathrm{clin}}(y_s, y_T^+) \le M_{\mathrm{clin}}(y_s, y_T^-), \\ \kappa_\tau^+ \varphi_\tau^{(+,0)}, & z = eg,\ M_{\mathrm{clin}}(y_s, y_T^+) \le M_{\mathrm{clin}}(y_s, y_T^0), \\ \kappa_\tau^+, & \text{otherwise}, \end{cases} \tag{2}$$

Since both curriculums use q_τ^+ as the final supervision target, their weighted OPSD objective can be uniformly written as Eq. 3:

$$\mathcal{L}_{\mathcal{C}_z} = \mathbb{E}_{y_s \sim \pi_S(\cdot|x)} \left[\frac{\sum_{\tau=1}^{|y_s|} w_\tau^z D_{\mathrm{KL}}(q_\tau^+ \parallel p_\tau)}{\sum_{\tau=1}^{|y_s|} w_\tau^z + \epsilon} \right]. \tag{3}$$

where ϵ is a numerical stabilization term. Through this rollout-adaptive curriculum distillation process, $\mathcal{C}_{ed}$ first strengthens the ability of the student policy to distinguish the diagnostic information associated with E_x^+ and E_x^-. Subsequently, $\mathcal{C}_{eg}$ further internalizes the incremental diagnostic information provided by E_x^+ into the student policy.

Reasoning Consolidation. In $\mathcal{C}_{rc}$, CA-OPSD removes E_x^+ and anneals the teacher privileged context to the GT report condition c^0. Since no external clinical evidence or evidence-condition comparison is introduced, curriculum-objective tendency evaluation is omitted, and only π_T^0 remains as the distillation target. Given $y_{s<\tau}$, the teacher distribution is q_τ^0, and token-level supervision is weighted by κ_τ^0. The corresponding OPSD objective is given by Eq. 4:

$$\mathcal{L}_{\mathcal{C}_{rc}} = \mathbb{E}_{y_s \sim \pi_S(\cdot|x)} \left[\frac{\sum_{\tau=1}^{|y_s|} \kappa_\tau^0 D_{\mathrm{KL}}(q_\tau^0 \parallel p_\tau)}{\sum_{\tau=1}^{|y_s|} \kappa_\tau^0 + \epsilon} \right]. \tag{4}$$

With clinical evidence removed, CA-OPSD reaches an evidence-free setting consistent with inference, enabling the optimized model π_θ^* to generate reports from the original radiological image while consolidating evidence-supported diagnostic reasoning into internal diagnostic reasoning capability.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Implementation Details: We adopt BLIP-3 (Xue et al., 2025) as the RRG model π_θ, with SigLIP as the visual encoder, Perceiver Resampler as the visual token sampler, and Phi-3-mini-3.8B as the language model. During supervised fine-tuning, all input images are resized to 384×384. The visual encoder and text decoder are fine-tuned using Low-Rank Adaptation (LoRA) with a rank of 64, while the Perceiver Resampler is fully fine-tuned. We optimize the RRG model using AdamW with an initial learning rate of 1×10^{-4} and a batch size of 96 for 5 epochs. All experiments are conducted on NVIDIA RTX PRO 6000 (96GB) GPUs. **More implementation details on multi-tool evidence orchestration, CA-OPSD, and the inference stage are provided in Appendix B.3**.

Datasets and Evaluation Metrics: Experiments are conducted on two public datasets, MIMIC-CXR and IU X-Ray. For MIMIC-CXR, we follow the official data split and perform data cleaning, while IU X-Ray follows the split adopted by R2Gen (Chen et al., 2020). For both datasets, only the findings section of each report is used as the generation target. During multi-tool evidence orchestration, we sample 160 training cases and 40 validation cases from CheXpert Plus (Chambon et al., 2024) for MCTS search and CSC-SkillEvo. For CA-OPSD, all training samples of IU X-Ray are used, while 3,000 cases are sampled from MIMIC-CXR for self-distillation. All sampling procedures consider the coverage of normal and abnormal cases, as well as common and rare findings. Following existing RRG evaluation protocols, we adopt BLEU-4 (BL-4), METEOR (MTR), and ROUGE-L (RG-L) to evaluate the linguistic quality of generated reports. At the clinical level, we adopt the Clinical Efficacy (CE) metric and report micro-averaged Precision (P), Recall (R), and F1-Score (F1) based on the CheXpert-14 labels to evaluate performance on the MIMIC-CXR dataset. For the smaller-scale IU X-Ray dataset, where disease distributions are relatively sparse, label-level CE metrics are susceptible to the effects of class sparsity and distribution imbalance. Therefore, we employ RadGraph F1 (Rad-F1) (Jain et al., 2021) and GREEN (Ostmeier et al., 2024) to assess clinical consistency.**For the evaluation of clinical evidence, we conducted additional measurement experiments in Appendix B.3.**

4.2 MAIN RESULTS

Table 1 compares the proposed method with representative RRG approaches on the MIMIC-CXR and IU X-Ray datasets, covering vision-language generation, LLM-based report generation, medical knowledge enhancement, and clinical semantic alignment. The results show that our method achieves the best overall performance on both datasets and demonstrates clear advantages in report generation quality. Notably, on MIMIC-CXR, our method shows a pronounced advantage in the CE metrics, indicating that the generated reports more accurately preserve key abnormalities and clinically relevant findings from the input images. These results further demonstrate that the proposed framework can extract high-quality clinical evidence and internalize the evidence-supported diagnostic reasoning into the RRG model, thereby improving the factual accuracy and clinical reliability of radiology report generation.

4.3 ABLATION STUDIES

Performance of different ablation components. Table 2 reports the contributions of the proposed components. For all settings without "CA-OPSD", including "MCTS" and "CSC-SkillEvo" variants, the clinical evidence obtained through search is first injected into the "RRG model" via off-policy distillation, followed by report SFT. Building "CSC-SkillEvo" upon "Base+MCTS" consistently improves both NLG and CE metrics, indicating that the $\mathbb{S}_{tp}$ and $\mathbb{S}_{st}$ agent skills enable "MCTS" to acquire higher-quality clinical evidence. Further applying "CA-OPSD" to the collected evidence yields additional gains in clinical metrics such as GREEN and Rad-F1, demonstrating the effectiveness of curriculum-annealed distillation in internalizing evidence-supported diagnostic rea-

Table 1: Comparison with state-of-the-art methods on the MIMIC-CXR and IU X-Ray.

Methods	Conferences	MIMIC-CXR						IU X-Ray		
		NLG Metrics			CE Metrics			NLG Metrics		
		BL-4	MTR	RG-L	P	R	F1	BL-4	MTR	RG-L
METransformer(Wang et al., 2023b)	CVPR 2023	0.124	0.152	0.291	0.364	0.309	0.311	0.172	0.192	0.380
ORGAN(Hou et al., 2023)	ACL 2023	0.123	0.162	0.293	0.416	0.418	0.385	0.195	0.205	0.399
SILC(Liu et al., 2024a)	TMI 2024	0.141	0.163	0.309	0.457	0.337	0.330	0.175	0.192	0.379
Bootstrapping(Liu et al., 2024b)	AAAI 2024	0.128	0.175	0.291	0.465	0.482	0.473	0.184	0.208	0.390
MambaXray(Wang et al., 2025a)	CVPR 2025	0.133	0.167	0.289	0.561	0.460	0.505	0.185	0.216	0.371
MLRG(Liu et al., 2025a)	CVPR 2025	0.158	0.176	0.320	0.549	0.468	0.505	–	–	–
DART(Park et al., 2025)	CVPR 2025	0.137	0.175	0.310	0.520	0.546	0.533	0.208	0.205	0.411
MultiP-R2Gen(Chen et al., 2025)	TMI 2025	0.140	0.167	0.307	–	–	–	0.178	0.216	0.386
KACL(Sha et al., 2025)	MICCAI 2025	0.136	0.169	0.303	0.503	0.442	0.469	0.184	0.211	0.385
SAT-RRG(Liu et al., 2026e)	CVPR 2026	0.143	0.167	0.303	–	–	–	0.196	0.222	0.400
PriorRG(Liu et al., 2026b)	AAAI 2026	0.175	0.189	0.324	0.541	0.485	0.511	–	–	–
Ours	**This Paper**	**0.224**	**0.223**	**0.355**	**0.604**	**0.554**	**0.572**	**0.267**	**0.239**	**0.428**

soning into the RRG model. Overall, these components work synergistically to establish a complete pipeline from clinical evidence search and orchestration to diagnostic reasoning internalization.

Table 2: Ablation study of different components on IU X-Ray. "Base" denotes the SFT RRG model.

Base	MCTS	CSC-SkillEvo		CA-OPSD	BL-4	MTR	RG-L	GREEN	Rad-F1
		$\mathbb{S}_{tp}$	$\mathbb{S}_{st}$						
✓					0.196	0.201	0.356	0.586	0.408
✓	✓				0.217	0.210	0.378	0.603	0.417
✓	✓	✓			0.241	0.220	0.397	0.612	0.429
✓	✓		✓		0.236	0.218	0.393	0.608	0.425
✓	✓	✓	✓		0.248	0.223	0.413	0.617	0.436
✓	✓	✓	✓	✓	**0.267**	**0.239**	**0.428**	**0.631**	**0.450**

Table 3: Comparison of different teacher token-weighting strategies across three curriculum stages on MIMIC-CXR.

Token Weight w_τ	BL-4	MTR	RG-L	P	R	F1
$w_\tau = 1$	0.199	0.191	0.337	0.568	0.521	0.549
$w_\tau = \kappa_\tau$	0.206	0.201	0.343	0.588	0.527	0.556
$w_\tau = \varphi_\tau$	0.211	0.208	0.349	0.592	0.532	0.564
Ours	**0.224**	**0.223**	**0.355**	**0.604**	**0.543**	**0.572**

Effect of different token weight strategies. Table 3 compares the impact of different token weight strategies on CA-OPSD. Specifically, $w = 1, w = \kappa_\tau$ and $w = \varphi_\tau$ denote uniform weighting, normalized-entropy confidence weighting, and teacher-discrepancy weighting for token-level supervision on student rollouts, respectively. The experimental results show that, compared with the uniform weighting strategy, introducing token confidence-based weighting enables the model to focus more on high-confidence tokens with diagnostic value, thereby effectively improving performance. Furthermore, privilege context-based token weighting leverages the prediction discrepancies among different teacher policies and dynamically adjusts the supervision intensity across different rollout stages, providing the student policy with more fine-grained and targeted learning signals. Finally, the Ours strategy, which integrates both types of token weighting, achieves the best performance, demonstrating that CA-OPSD can effectively utilize external clinical evidence and progressively transfer and internalize it into the intrinsic reasoning capability of the student policy. **More extensive quantitative results, qualitative analyses, and case studies are provided in Appendices C.1, C.2, and C.3, respectively**.

5 Conclusion

We propose an evidence-informed RRG framework that aims to implicitly integrate evidence-supported reasoning into the report generation process. During deliberative multi-tool evidence orchestration, we first incorporate CSC-SkillEvo to optimize evidence search and facilitate the coordinated integration of evidence from multiple tools. CA-OPSD further adopts progressive curriculum learning, enabling the RRG model to develop stable clinical reasoning capabilities as access to evidence is progressively reduced. Extensive experiments demonstrate that the proposed method can transform external medical cues into intrinsic clinical reasoning capabilities within the RRG model.

AI USE STATEMENT

During the preparation of this manuscript, we used large language models (LLMs) to assist with language-related tasks only. Specifically, LLMs were used for language polishing, grammar and spelling checking, terminology consistency, and improving clarity and coherence of expressions. All LLM-assisted content was carefully reviewed, verified, and revised by the authors as necessary. All scientific contributions of this work, including the research ideas, methodology, analyses, and conclusions, were developed and conducted by the authors independently. The authors take full responsibility for the content of this manuscript, including all claims, interpretations, and the final version submitted.

ETHICS STATEMENT

We acknowledge that we and all co-authors of this work have read and commit to adhering to the ICLR Code of Ethics. We use de-identified medical imaging data approved for research purposes. As automatically generated medical reports may contain errors or biases, our method is intended for research only and should not be used for clinical decision-making without professional review.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed method, CSC-SkillEvo, CA-OPSD, and the experimental setup in the main paper and appendix. Implementation details, including data processing, model configurations, training hyperparameters, tool settings, and evaluation protocols, are provided in the appendix.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Rina Bao, Shilong Dong, Zhenfang Chen, Sheng He, Ellen Grant, and Yangming Ou. Visual and domain knowledge for professional-level graph-of-thought medical reasoning. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 2930–2952. PMLR, 13–19 Jul 2025.

Walid Bousselham, Hilde Kuehne, and Cordelia Schmid. Vold: Reasoning transfer from llms to vision-language models via on-policy distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26209–26218, June 2026.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P. Langlotz. Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats, 2024.

Haoxuan Che, Haibo Jin, Zhengrui Guo, Yi Lin, Cheng Jin, and Hao Chen. Llm-driven medical report generation via communication-efficient heterogeneous federated learning. *IEEE Transactions on Medical Imaging*, 45(1):28–39, 2026. doi: 10.1109/TMI.2025.3591185.

Zailong Chen, Yingshu Li, Zhanyu Wang, Peng Gao, Johan Barthelemy, Luping Zhou, and Lei Wang. Enhancing radiology report generation via multi-phased supervision. *IEEE Transactions on Medical Imaging*, 44(11):4666–4677, 2025. doi: 10.1109/TMI.2025.3580659.

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 (eds.),

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1439–1449, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.112.

Zhuoxiao Chen, Hongyang Yu, Ying Xu, Yadan Luo, Long Duong, and Yuan-Fang Li. Orapo: Oracle-educated reinforcement learning for data-efficient and factual radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28275–28287, June 2026.

Rémi Coulom. Efficient selectivity and backup operators in monte-carlo tree search. In H. Jaap van den Herik, Paolo Ciancarini, and H. H. L. M. (Jeroen) Donkers (eds.), *Computers and Games*, pp. 72–83, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-75538-8.

Chengwei Dai, Kun Li, Wei Zhou, and Songlin Hu. Capture the key in reasoning to enhance CoT distillation generalization. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 441–465, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.21.

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.

Yushi Feng, Junye Du, Yingying Hong, Qifan Wang, and Lequan Yu. Pass: Probabilistic agentic supernet sampling for interpretable and adaptive chest x-ray reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(36):30717–30725, 2026. doi: 10.1609/aaai.v40i36.40328.

Fatemeh Ghezloo, Mehmet Saygin Seyfioglu, Rustin Soraki, Wisdom O. Ikezogwo, Beibin Li, Tejoram Vivekanandan, Joann G. Elmore, Ranjay Krishna, and Linda Shapiro. Pathfinder: A multi-modal multi-agent system for medical diagnostic decision-making applied to histopathology. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23431–23441, October 2025.

Wenjun Hou, Kaishuai Xu, Yi Cheng, Wenjie Li, and Jiang Liu. ORGAN: Observation-guided radiology report generation via tree reasoning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8108–8122, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.451.

ZhiYan Hou, Xinyu Tang, Hongyan An, Jianjin Zhang, Weizhen Wang, Yunyun Han, Gengsheng Li, Xiangzhao Hao, Haiyun Guo, Wenbin Hu, Jinqiao Wang, and Yafeng Deng. Dash: Divergence-adaptive supervision horizons for on-policy self-distillation of reasoning models, 2026.

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 (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Baoyu Jing, Zeya Wang, and Eric Xing. Show, describe and conclude: On exploiting the structure information of chest X-ray reports. In Anna Korhonen, David Traum, and Llu
'is M
àrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6570–6580, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1657.

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, 2019. doi: 10.1038/s41597-019-0322-0.

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 (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Aohan Liu, Yuchen Guo, Jun-Hai Yong, and Feng Xu. Multi-grained radiology report generation with sentence-level image-language contrastive learning. *IEEE Transactions on Medical Imaging*, 43(7):2657–2669, 2024a. doi: 10.1109/TMI.2024.3372638.

Chang Liu, Yuan 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, 2024b. doi: 10.1609/aaai.v38i17.29826.

Gang Liu, Xiaotian Tang, Jiacheng Gan, Tianyu Ren, Tingyao Liu, and Shenjun Zhong. *MMG-RAG: Multimodal Graph RAG for Medical Report Generation*, pp. 2477–2485. Association for Computing Machinery, New York, NY, USA, 2026a. ISBN 9798400726170.

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)*, pp. 10348–10359, June 2025a.

Kang Liu, Ziyang Ma, Zhen Fang, Yuhong Li, Kexin 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, 2026b. doi: 10.1609/aaai.v40i9.37657.

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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5595–5603, 2025b. doi: 10.1609/aaai.v39i6.32596.

Tengfei Liu, Yijian Fan, Boyue Wang, Yongli Hu, Mingjie Li, Jinghua Li, Junbin Gao, Xiaojun Chang, Zhihui Li, and Baocai Yin. Biotprompt: Bidirectional optimal transport guided prompting for disease evolution-aware radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13755–13765, June 2026c.

Xiaogeng Liu, Xinyan Wang, Yingzi Ma, Yechao Zhang, and Chaowei Xiao. When are teacher tokens reliable? position-weighted on-policy self-distillation for reasoning. *CoRR*, abs/2605.21606, May 2026d.

Yunyi Liu, Yingshu Li, Tong Chen, Lingqiao Liu, Lei Wang, and Luping Zhou. Sat-rrg: Llm-guided self-adaptive training for radiology report generation with token-level push-pull optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 35363–35372, June 2026e.

Yuxing Lu, Gecheng Fu, Wei Wu, Xukai Zhao, Sin Yee Goi, and Jinzhuo Wang. Towards doctor-like reasoning: Medical rag fusing knowledge with patient analogy through textual gradients. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 150336–150375. Curran Associates, Inc., 2025. doi: 10.52202/085713-5027.

Xuan-Phi Nguyen, Zeyu Leo Liu, Yang Li, Shrey Pandit, Yiran Zhao, Anurag Koul, and Shafiq Joty. Privileged likelihood is not automatically value: Three checks for token credit in on-policy self-distillation, 2026.

Andrei Liviu Nicolicioiu, Mohammad Pezeshki, and Aaron Courville. On-policy self-distillation with sampled demonstrations reduces output diversity. In *ICML 2026 Workshop on Foundations of Deep Generative Models: Understanding Memorization, Generalization, and Reasoning*, 2026.

Timothy Ossowski, Sheng Zhang, Qianchu Liu, Guanghui Qin, Reuben Tan, Tristan Naumann, Junjie Hu, and Hoifung Poon. Octomed: Data recipes for state-of-the-art multimodal medical reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26251–26261, June 2026.

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 (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 374–390, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.21.

Sang-Jun Park, Keun-Soo Heo, Dong-Hee Shin, Young-Han Son, Ji-Hye Oh, and Tae-Eui Kam. Dart: Disease-aware image-text alignment and self-correcting re-alignment for trustworthy radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15580–15589, June 2025.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, dahai li, Zhiyuan Liu, and Maosong Sun. Toolllm: Facilitating large language models to master 16000+ real-world apis. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 9695–9717, 2024.

Melanie Rieff, Maya Varma, Ossian Rabow, Subathra Adithan, Julie Kim, Ken Chang, Hannah Lee, Nidhi Rohatgi, Christian Bluethgen, Mohamed Muneer, Jean-Benoit Delbrouck, and Michael Moor. Smmile: An expert-driven benchmark for multimodal medical in-context learning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-2180.

Christopher D. Rosin. Multi-armed bandits with episode context. *Annals of Mathematics and Artificial Intelligence*, 61(3):203–230, 2011. ISSN 1573-7470. doi: 10.1007/s10472-011-9258-6.

Yuyang Sha, Hongxin Pan, Weiyu Meng, and Kefeng Li. Contrastive Knowledge-Guided Large Language Models for Medical Report Generation . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume LNCS 15965. Springer Nature Switzerland, September 2025.

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, 2024.

Guobin Shen, Lei Huang, Xiang Cheng, Chenxiao Zhao, Jindong Li, Dongcheng Zhao, and Xing Yu. From generic correlation to input-specific credit in on-policy self distillation, 2026.

Samyak Shrestha and Alexander Tessier. Rethinking privileged information in on-policy self-distillation. *arXiv preprint arXiv:2608.18271*, 2026.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2609–2634, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.147.

Xiao Wang, Fuling Wang, Yuehang Li, Qingchuan Ma, Shiao Wang, Bo Jiang, and Jin Tang. Cxpmrg-bench: Pre-training and benchmarking for x-ray medical report generation on chexpert plus dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5123–5133, June 2025a.

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)*, pp. 11558–11567, June 2023b.

Zhuhao Wang, Yihua Sun, Zihan Li, Xuan Yang, Fang Chen, and Hongen Liao. Llm-rg4: Flexible and factual radiology report generation across diverse input contexts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8250–8258, 2025b. doi: 10.1609/aaai.v39i8.32890.

Ziao Wang, Sixing Yan, Kejing Yin, Xiaofeng Zhang, and William K. Cheung. Curv: Coherent uncertainty-aware reasoning in vision-language models for x-ray report generation. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 93686–93713. Curran Associates, Inc., 2025c. doi: 10.52202/085713-3133.

Shuyao Xu, Cheng Peng, Jiangxuan Long, Weidi Xu, Wei Chu, and Yuan Qi. Harnessing negative signals: Reinforcement distillation from teacher data for LLM reasoning. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1618–1639, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.74.

Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, Shrikant Kendre, Jieyu Zhang, Shaoyen Tseng, Gustavo Adolfo Lujan-Moreno, Matthew Lyle Olson, Musashi Hinck, David Cobbley, Vasudev Lal, Can Qin, Shu Zhang, Chia-Chih Chen, Ning Yu, Juntao Tan, Tulika Manoj Awalgaonkar, Shelby Heinecke, Huan Wang, Yejin Choi, Ludwig Schmidt, Zeyuan Chen, Silvio Savarese, Juan Carlos Niebles, Caiming Xiong, and Ran Xu. Blip-3: A family of open large multimodal models, October 2025.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 11809–11822. Curran Associates, Inc., 2023a. doi: 10.52202/075280-0517.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, Kigali, Rwanda, 2023b.

Hee Suk Yoon, Eunseop Yoon, Jaehyun Jang, SooHwan Eom, Ji Woo Hong, Mark A. Hasegawa-Johnson, Qi Dai, Chong Luo, and Chang D. Yoo. Decomposed on-policy distillation for vision-language reasoning: Steering gradients for visual grounding. In *Forty-third International Conference on Machine Learning*, 2026.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. In *Forty-third International Conference on Machine Learning*, 2026a.

Ziqi Zhao, Xinyu Ma, Liu Yang, Yujie Feng, Daiting Shi, Jingzhou He, Xin Xin, Zhaochun Ren, and Xiao-Ming Wu. Rosd: Reflective on-policy self-distillation for language model reasoning across domains, 2026b.

Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language agent tree search unifies reasoning, acting, and planning in language models. In *Forty-first International Conference on Machine Learning*, 2024.

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)*, pp. 22529–22538, October 2025.

Qingqing Zhu, Tejas Sudharshan Mathai, Pritam Mukherjee, Yifan Peng, Ronald M. Summers, and Zhiyong Lu. Utilizing longitudinal chest x-rays and reports to pre-fill radiology reports. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pp. 189–198, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-43904-9.

Xinlin Zhuang, Feilong Tang, Haolin Yang, Xiwei Liu, Ming Hu, Huifa Li, Haochen Xue, Junjun He, Zongyuan Ge, Yichen Li, Ying Qian, and Imran Razzak. Towards efficient medical reasoning with minimal fine-tuning data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20223–20232, June 2026.

APPENDIX

A Detailed Algorithm Workflow 16
B Implementation Details 16
B.1 Radiology Tool Library and Interface Schemas 16
B.2 Prompt Template 17
B.3 Additional Training and Inference Details 23
C Experimental Details 24
C.1 Additional Quantitative Results 24
C.1.1 Evidence Quality Analysis 24
C.1.2 Robustness and Generalization Analysis 25
C.1.3 Reasoning Internalization Analysis 26
C.2 Additional Qualitative Results 27
C.3 Case Studies 30
D Limitations and Discussion 33

A Detailed Algorithm Workflow

Algorithm 1 summarizes the overall training workflow of the proposed framework, which consists of two stages, namely deliberative multi-tool evidence orchestration and CA-OPSD. Notably, CSC-SkillEvo, an agent skill self-evolution module, is integrated into the first stage.

Algorithm 1 Overall Training Pipeline

Require: Image–report pairs (x, y^*); GT-supervised π_θ; Decision Agent $\mathcal{A}_\theta$, Skill Optimizer $\mathcal{A}_s$; tools $\mathbb{T}_{\text{rad}}$; initial skills $\mathbb{S}_{tp}^0, \mathbb{S}_{st}^0$.
Ensure: Optimized RRG model π_θ^*.
▷ ***Stage 1: Deliberative Multi-tool Evidence Orchestration***
Initialize MCTS and CSC-SkillEvo with $\mathbb{S}_{tp}^0$ and $\mathbb{S}_{st}^0$
for each CSC-SkillEvo evolution round r **do**
  for each search sample (x, y^*) **do**
    Initialize s_0 from the anchor findings identified by Classifier
    while budget remains and the frontier is expandable **do**
      Select a_t^* by PUCT according to Eq. 1
      Expand Top3 actions from $\Upsilon(s_t)$ using $P(a \mid s_t)$ and $\mathbb{S}_{tp}^r$
      Execute a_t^*; obtain the resulting observation and E; update s_{t+1}
      Evaluate $V(s_{t+1})$, apply $\mathbb{S}_{st}^r$, and backpropagate N, W, Q
    end while
  end for
  Construct $\mathcal{D}_{tp}^r$ and $\mathcal{D}_{st}^r$ from MCTS comparisons
  Generate $\widetilde{\mathbb{S}}_{tp}^r$ and $\widetilde{\mathbb{S}}_{st}^r$ using $\mathcal{A}_s$
  Update $\mathbb{S}_{tp}^{r+1}$ and $\mathbb{S}_{st}^{r+1}$ after the validation gate
end for
Obtain positive and negative evidence trajectories E_x^+ and E_x^- using evolved skills and MCTS
▷ ***Stage 2: Curriculum-Annealed On-Policy Self-Distillation***
Set $c^+ = (y^*, E_x^+)$, $c^- = (y^*, E_x^-)$, and $c^0 = y^*$
Define curriculum stages $\mathcal{C} = (\mathcal{C}_{ed}, \mathcal{C}_{eg}, \mathcal{C}_{rc})$ for discrimination, grounding, and consolidation
Initialize $\pi_S, \pi_T^+, \pi_T^-, \pi_T^0 \leftarrow \pi_\theta$
for each curriculum stage $\mathcal{C}_k \in \mathcal{C}$ sequentially and each mini-batch **do**
  Sample $y_s \sim \pi_S(\cdot \mid x)$ and compute $p_\tau = \pi_S(\cdot \mid x, y_{s<\tau})$
  Select active teacher contexts $\alpha \in \{+, -\}, \{+, 0\}, \{0\}$ for $\mathcal{C}_{ed}, \mathcal{C}_{eg}, \mathcal{C}_{rc}$, respectively
  Evaluate teacher distribution

$$q_\tau^\alpha = \pi_T^\alpha(\cdot \mid x, c^\alpha, y_{s<\tau})$$

  if $\mathcal{C}_k \in \{\mathcal{C}_{ed}, \mathcal{C}_{eg}\}$ **then**
    Generate teacher rollout $y_T^\alpha \sim \pi_T^\alpha(\cdot \mid x, c^\alpha)$
    Compute token-level weights w_τ^z with M_{clin} according to Eq. 2, and curriculum loss $\mathcal{L}_{\mathcal{C}_z}$ according to Eq. 3
  else
    Compute κ_τ^0 and $\mathcal{L}_{\mathcal{C}_{rc}}$ according to Eq. 4
  end if
  Update only student parameters θ using curriculum loss while keeping all teachers frozen
end for
return Optimized model π_θ^*

B Implementation Details

B.1 Radiology Tool Library and Interface Schemas

To support multi-tool clinical evidence acquisition, we construct a radiology tool library $\mathbb{T}_{\text{rad}}$ comprising seven off-the-shelf radiology models and auxiliary tools. Each tool is exposed through a unified interface schema that specifies its functionality and multimodal input–output formats, as summarized in Figure 2. According to their functionality, these tools are organized into three categories: global diagnostic tools, localized analysis tools, and finding-level verification tools.

Global Diagnostic Tools. This category provides global diagnostic cues and initializes subsequent evidence search, including **Classifier** and **Retrieval-Augmented Generation (RAG)**. Given a radiology image, **Classifier** performs multi-label classification to identify potential pathological

findings together with their confidence scores. **RAG** takes the radiology image as the retrieval query, searches for visually similar historical cases, and returns the corresponding radiology reports as external clinical references.

Localized Analysis Tools. This category performs localized analysis around a target finding, including **Grounding, Segment**, and **Bone Suppression**. Given a radiology image and a target finding, **Grounding** localizes the corresponding candidate region and returns its bounding-box coordinates. Based on the radiology image and the resulting bounding box, **Segment** further generates a segmentation mask and an overlay view of the target region. **Bone Suppression** generates a bone-suppressed view from the given radiology image and localized region, facilitating the observation of abnormalities that may be obscured by bony structures.

Finding-Level Verification Tools. This category is used to verify target findings and includes **Visual Question Answering (VQA)** and **Gradient-weighted Class Activation Mapping (Grad-CAM)**. **VQA** takes the radiology image and a finding-specific visual question as input and uses an off-the-shelf medical VLM to produce the corresponding textual answer. **Grad-CAM** takes the radiology image and target finding as input to generate the corresponding activation region, which is then compared with the existing bounding box or segmentation mask to assess its consistency with localized evidence.

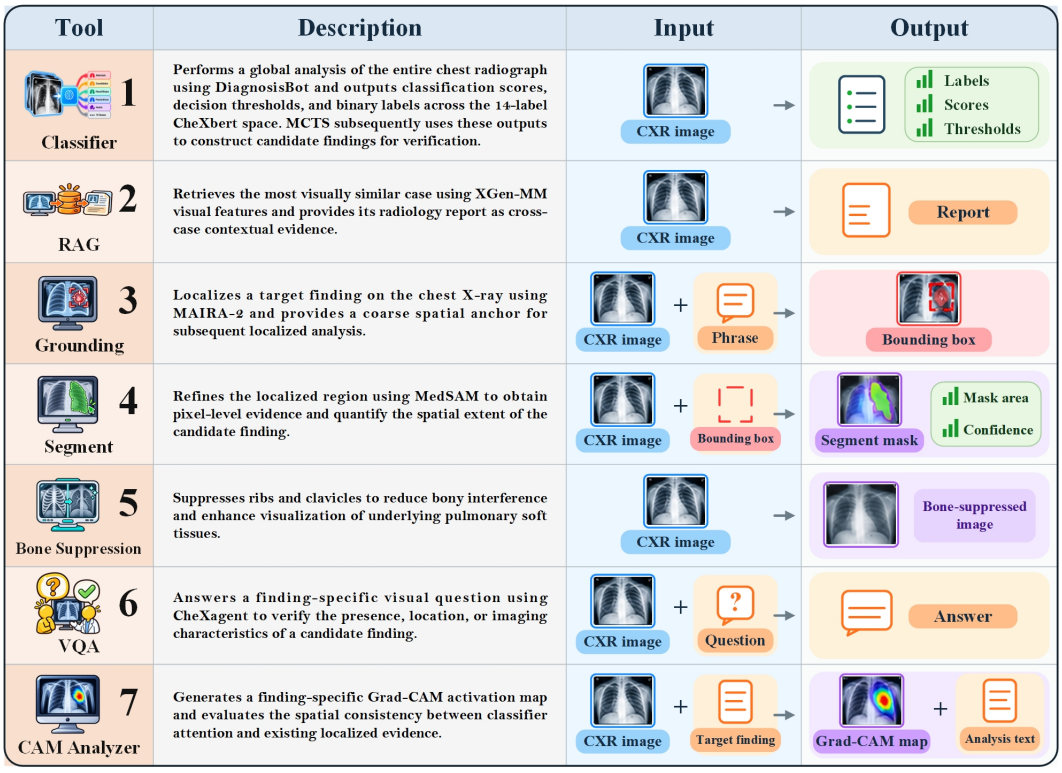

Figure 2: Overview of the radiology tool library $\mathbb{T}_{\text{rad}}$ and its multimodal interface schemas. Each tool is characterized by its functionality and input-output modalities.

B.2 Prompt Template

This section summarizes the prompt templates employed in the proposed framework and organizes them into the following two categories based on their functional roles:

(1) Prompt Design for Evidence Orchestration.

These prompts support tool decision-making, search-termination decisions, and evidence-context integration during multi-tool search. The tool-prioritization prompt evaluates the priorities of candidate tool actions at the same node based on the current search state and the current skill rules,

thereby guiding subsequent search expansion, as shown in Figure 3. The search-termination prompt determines whether to continue or terminate the search according to the current accumulated evidence state, as shown in Figure 4. The summary prompt integrates heterogeneous observations returned by different tools together with the corresponding search context, producing structured clinical evidence to update the current evidence state, as shown in Figure 5.

Tool-Prioritization Prompts

System Prompt:

At the current MCTS search path, score one or more candidate tool actions before execution.
For each candidate, assign the requested continuous priority score in [0, 1] by applying the current tool-prioritization skill holistically; higher scores indicate greater priority for execution.

When sibling candidates are evaluated jointly, calibrate their scores on a shared scale.
Base each judgment on the target's tool-use history along the current path and the candidate's exact current tool call. Do not infer unobserved tool outcomes. Follow the requested output schema exactly and provide one concise scoring rationale per candidate.

User Prompt:

Current Evidence State
<branch summary>

Candidate Tool Actions
[
 {
   "candidate_id": "<candidate id>",
    "target_scope": <target scope>,
    "classifier_confidence": <classifier confidence>,
     "current_context": {"target_history": <target history>},
      "candidate_tool_action": <candidate tool action>
       },
  ...
]

Task
Score every sibling candidate under the active Skill rules and provide one brief
reason for each overall assessment.
For each candidate, use only its corresponding `target_scope`, `classifier_confidence`,
`current_context.target_history`, and `candidate_tool_action`.

Use the Branch Evidence
Summary only for cross-target coherence.Do not infer unobserved tool outcomes. Calibrate the requested scores on a shared scale across the sibling set.

Output Contract
{
     "scores": {
     "<candidate id>": {
     "components": {
     "priority_score": "<score 0 to 1>"},
      "reason": "<brief reason>"}}
}

Figure 3: Design details for tool-prioritization prompt.

Search-Termination Prompts

System Prompt:

At the current MCTS search path, decide whether the current clinical evidence is sufficient under
the current Search-Termination Skill
Base the decision only on Current Clinical Evidence and Tool-Use History supplied in the request.

Return exactly {"stop":true} or {"stop":false}.

User Prompt:

Anchor Findings
[
"<finding id>",
...
]

Current Clinical Evidence
{"clinical_evidence": "<current clinical evidence>"}

Tool-Use History
[
{"tool": <tool>,
"args": <args>,
"observation": <observation>,
"finding": <finding>},
...
]

Task
Decide whether the current clinical evidence is sufficient without another tool action.

Output Contract
{"stop": "<boolean>"}

Figure 4: Design details for search-termination prompt.

Summery Prompts

System Prompt:

```
Revise the parent structured clinical evidence by integrating exactly one
newly executed
observation. Treat the parent clinical evidence as an editable draft.

Add, remove, negate, qualify, or merge evidence statements wherever the
new output changes
the best evidence-supported account; express unresolved conflicts as
calibrated uncertainty.
Ground every revision in the actual output: finding labels and arguments are
queries, and
RAG-retrieved radiology reports are indirect clinical context.
Produce the shortest complete structured clinical evidence paragraph,
omitting tool names,
process narration, rationales, and repetition.

Do not use classifier priors, prevalence, or ground truth.
If no current-image finding is supported, state that the findings remain
indeterminate.
Return exactly
{"findings":"<one concise plain paragraph>"}.
```

User Prompt:

```
## Current Context
 {
    "parent_clinical_evidence": "<parent clinical evidence>"
    "new_observation": {
    "tool": "<tool>",
    "args": <args>,}
 }

## Task
Revise parent_clinical_evidence to fully integrate new_observation.
Update every affected claim  and return the shortest complete structured
clinical evidence paragraph.

## Output Contract
{"clinical_evidence":"<one concise plain paragraph>"}
```

Figure 5: Design details for evidence summarization prompt.

(2) Prompt Design for CSC-SkillEvo.

These prompts enable the continual refinement of agent skills based on feedback from MCTS. The diagnosis prompt compares decision outcomes across counterfactual samples to identify deficiencies in the current skill while extracting effective decision behaviors that should be preserved, as shown in Figure 6. The update intention prompt generates structured skill-update intentions from the diagnosis results and specifies the directions for rule adjustment, as shown in Figure 7. The revision prompt holistically revises the existing skill rules according to the update intentions and produces a candidate skill, thereby enabling continual refinement of tool-selection and search-control strategies, as shown in Figure 8.

Diagnosis Prompts

System Prompt:

You are the Diagnosis Agent for text-Skill optimization.

Each example compares a host-selected preferred decision with a contrasted decision in the same context. Accept that ordering as fixed. Decisions may be sibling evidence-tool actions or STOP versus an executed one-step continuation. Follow the task description in the user message. The active Skill contradicted every supplied counterfactual feedback. Diagnose the reusable rule defect that should be revised.

Infer only the strongest recurring, atomic, decision-time policy diagnosis in the minibatch. Keep its observable trigger, preferred decision, contrasted decision, and one supported exception separate. Prefer the most concrete wording justified by more than one example: retain a recurring finding name, its evidence state, and the actual tool names.

If finding names differ, use a narrow shared clinical class only when the same decision direction genuinely holds; do not invent a shared scope.

For tool prioritization, describe both tools and their target or purpose whenever those details recur. For search termination, describe STOP versus CONTINUE and the feasible next tool whenever the examples support it.

Avoid replacing available evidence details with generic phrases. A generic actionable diagnosis is less useful than an honest abstention when the minibatch has opposing tool directions or no shared scope.

For STOP, the trigger must use only supplied decision-time context. Observed effects may justify the ordering but must not become a hindsight-only trigger.
Keep the rationale to one brief evidence-grounded sentence. Do not infer unobserved outcomes or write the final Skill. Exclude identifiers, hidden reports, evaluators, rewards, and numeric thresholds.

Return only the following JSON object:

```
{
    "condition": "<condition>",
    "preferred_decision": "<preferred decision>",
    "contrasted_decision": "<contrasted decision>",
    "exception": "<supported exception>",
    "rationale": "<brief rationale>"
}
```

Set exception to null when no supported exception applies.
If the minibatch does not support one coherent reusable atomic boundary, return condition, preferred_decision, contrasted_decision, and exception as null, plus a brief rationale. Use short phrases, not a multi-rule paragraph.

User Prompt:

Counterfactual contexts

```
{
    "task": {
    "task": "<task>",
    "target_skill": "<target skill>",
    "decision": "<decision>"
        },
        "policy_boundary": "<policy boundary>",
        "skill": "<target skill body>",
        "pairs": [{
        "preferred": <preferred decision>,
        "contrasted": <contrasted decision>,
        "context": <context>
        }]
}
```

Figure 6: Prompt template for diagnosis.

Update Intention Prompts

System Prompt:

You are the Skill Update Intention Agent.

Given one target Skill and concise, boundary-keyed Revise and Keep diagnosis logs,
propose up to the allowed number of independent evidence-grounded rule edits.
Correction evidence identifies a defect; Preservation evidence supplies a guard or exception for the same policy boundary. Do not combine unrelated boundaries into one proposal.

Use add for a missing standalone rule, delete for removing redundant or conflicting rules, and revise for modifying an existing rule's condition, scope, or applicability.
When the Skill is below its target rule count, prefer supported add proposals over unnecessarily merging signals into existing generic rules.

Each proposal must cite supplied diagnosis IDs from exactly one boundary, listing a Revise diagnosis first and any supporting Keep diagnosis after it.
Name the target rule when the operation edits an existing rule.

Keep proposals independent and rank the strongest first.
Prefer instructions that preserve the observable condition, preferred and contrasted decision, concrete medical/tool scope, and supported exception.

Avoid generic phrases such as "different finding", "complementary tool", or "targeted follow-up" when the actual finding and tool are available.
Do not write the revised Skill.

Return only the following JSON object:

```
{
  "proposals": [{
      "boundary_key": "<boundary key>",
      "operation": "<operation>",
      "target_rule": <target rule or null>,
      "instruction": "<instruction>",
      "evidence_diagnosis_ids": <evidence diagnosis ids>} ]
  }
```

User Prompt:

Diagnosis logs

```
{
    "task": {
    "task": "<task>",
    "target_skill": "<target skill>",
    "decision": "<decision>"
 },
    "skill": "<current skill body>",
    "rule_capacity": {
    "current": <current rule count>,
    "add_allowed": <add allowed>,
    "target": <target rule count>
    },
    "diagnoses": [{
        "diagnosis_id": "<diagnosis id>",
        "policy_boundary": "<policy boundary>",
        "evidence_count": <evidence count>,
        "role": "<role>",
        "condition": "<condition>",
        "preferred_decision": "<preferred decision>",
        "contrasted_decision": "<contrasted decision>",
        "exception": <exception>,
        "rationale": "<rationale>"}]
}
```

Figure 7: Design details for update intention.

Revision Prompts

System Prompt:

You are the Skill Revision Agent for one proposed stage-scoped Skill update.

Apply every supplied non-conflicting operation to the current Skill and return its complete revised body.

Keep between two and five numbered rules under "## Scoring Rules", use no third-level headings, and keep every rule concise and atomic.

Tool-Prioritization Skill rules govern tool priority; Search-Termination Skill rules govern STOP versus CONTINUE decisions.

Each add appends one standalone rule. Each revise changes its named rule without changing the count. Each merge combines redundant rules and may reduce the count by one.

Apply at most two operations. Do not hide new rules inside semicolon clauses.
Preserve unaffected behavior.Prefer learned rules that retain the observable medical condition, concrete preferred and contrasted tool/decision, and supported exception instead of replacing them with generic
informativeness language.

When the update evidence supplies the detail, write the actual finding or narrow class, its evidence state, the exact tool names, and their ordering.
For stopping, make STOP versus CONTINUE and the proposed next tool explicit.

Follow the supplied tool capabilities:
VQA tools test presence or one attribute.
Grounding tools provide spatial localization evidence.
Segment, Grad-CAM, and Bone Suppression provide localized analysis refinements.
RAG provides semantic retrieval context rather than primary visual evidence.
Avoid placeholders such as "another finding", "targeted follow-up", or "complementary tool" when concrete terms are available.

Do not invent medical specificity absent from the diagnoses.
Do not restate host-owned execution or output constraints.

Return only the following JSON object:
{"skill": "<complete revised skill body>"}

If the proposed update cannot be applied without violating this contract, return {"skill": null}.

User Prompt:

Update intentions
```
{
    "task": {
    "task": "<task>",
    "target_skill": "<target skill>",
    "decision": "<decision>"},
    "skill": "<current skill body>",
    "update": {
    "operations": [{
                "target_rule": <target rule>,
                "instruction": "<instruction>"}]},
    "rule_contract": {
                 "current_count": <current count>,
                "expected_count": [<expected counts>],
                 "minimum": 2,
                "maximum": 5}
}
```

Figure 8: Design details for revision prompt.

B.3 Additional Training and Inference Details

This section provides further implementation details of our framework during training and inference, including the configurations of MCTS, CSC-SkillEvo, and CA-OPSD, as well as the evaluation settings for report generation and clinical evidence at test time.

Training Details. During multi-tool evidence orchestration, we employ Qwen3.5-9B as the decision-making agent $\mathcal{A}_\theta$. During search, the expansion stage retains the Top3 candidate actions with priority scores $\sigma > 0.5$. For each sample, MCTS performs at most 30 simulations with a maximum search depth of 8, and the exploration coefficient $\tilde{c}$ in PUCT is set to 1.5. For CSC-SkillEvo, we use 160 training samples for agent skill optimizer $\mathcal{A}_s$, with a batch size of 32 over 3 epochs. GPT-5.5 is adopted as the Skill optimizer, with the reasoning effort set to High.

For CA-OPSD training, we sample 8 student on-policy rollouts per sample with a temperature of 1.0, and set the maximum generation length of each rollout to 256 tokens. The curriculum stages are trained sequentially in the order of $C_{\rceil\lceil}$, $C_{\rceil\}}$, and $C_{\nabla\rfloor}$, with 30 optimization steps performed at each stage. To improve training stability, the token-level KL divergence is clipped to $[0, 0.05]$. The model is optimized using AdamW with a batch size of 32 and a learning rate of 5×10^{-6}.

For constructing the hard-negative evidence E_x^-, we first use CheXbert to extract CheXpert-14 labels from both the reference report and each candidate clinical evidence, and compute the corresponding F1 score. We then define $1 - F1$ as the label error rate. This error rate is combined with the BLEU-4 score by taking their harmonic mean, and the candidate trajectory with the highest value is selected as the hard-negative clinical evidence E_x^-. In contrast, the clinical evidence derived from the optimal MCTS trajectory is used as the positive evidence E_x^+. These positive and negative evidence pairs are further used to construct the privileged contexts for different curriculum stages in CA-OPSD.

Inference Details. At test time, the model no longer relies on external clinical evidence or any external tools. Only the student model trained with CA-OPSD is retained, which directly generates radiology reports from the input images. Therefore, neither MCTS nor additional tool calls are required during inference. Report generation is performed using beam search with a maximum generation length of 256 tokens, a length penalty of 1.5, and a beam size of 5.

To evaluate the quality of clinical evidence, we adopt three types of metrics, including Evi-F1, CQ_{LLM}, and CQ_{Rad}. For Evi-F1, CheXbert is used to extract CheXpert-14 labels from both the reference report and the generated clinical evidence, and their F1 score is computed to measure disease-level consistency with the reference. For CQ_{LLM}, GPT-5.5 evaluates the generated clinical evidence against the reference report in terms of image faithfulness and reasoning quality, and the final score is normalized to $[0, 1]$. The reasoning effort is set to High during evaluation. For CQ_{Rad}, five radiologists independently assess the generated clinical evidence according to the imaging findings documented in the reference report, focusing on report faithfulness and clinical accuracy. Each evaluation dimension is rated using a four-level scale of $[0, 0.3, 0.6, 1.0]$.

C Experimental Details

C.1 Additional Quantitative Results

In this section, we provide a comprehensive analysis of clinical evidence by combining objective automatic evaluation with subjective quality assessment. Specifically, Evi-F1 serves as an objective metric based on CheXbert, while CQ_{LLM} and CQ_{Rad} measure subjective clinical quality using large language models and expert radiologist evaluations, respectively. Detailed computation procedures are provided in Appendix B.3.

C.1.1 Evidence Quality Analysis

Comparison of Different Tool-Augmented Decision-making. As shown in Table 4, the two reasoning-guided search methods incur relatively low search overhead but are limited by the lack of explicit multi-path exploration, resulting in weaker evidence quality and clinical reasoning performance. Heuristic search methods introduce additional search cost for modest gains in Evi-F1, CQ_{Rad}, and CQ_{LLM}. In contrast, tree-structured search explores richer evidence trajectories through multi-branch search, improving both evidence acquisition and reasoning quality with only a limited increase in search cost. Among these methods, CSC-SkillEvo achieves the best overall performance, showing that self-evolving skill priors can effectively guide evidence selection in MCTS while maintaining a reasonable search budget.

Table 4: Comparison of evidence quality and tool-use efficiency across different tool-augmented decision-making methods on 100 cases from MIMIC-CXR and IU X-Ray. "Tool Calls" denotes the average number of tool invocations required by each method along its optimal search trajectory. "*" denotes methods re-implemented from the original papers for clinical evidence extraction.

Method	Evi-F1	CQ_{LLM}	CQ_{Rad}	Tool-Calls
Reasoning-guided Search				
ReAct* (Yao et al., 2023b)	0.476	0.461	0.531	5.25
Plan-and-Solve* (Wang et al., 2023a)	0.461	0.386	0.494	5.34
Heuristic Search				
Greedy Search	0.414	0.480	0.418	5.59
Beam Search	0.547	0.550	0.662	7.91
Tree-structured Search				
DFSDT* (Qin et al., 2024)	0.652	0.618	0.721	6.91
ToT* (Yao et al., 2023a)	0.542	0.467	0.659	6.85
MCTS	0.606	0.630	0.692	**4.77**
CSC-SkillEvo (Ours)	**0.723**	**0.719**	**0.846**	5.27

Effect of Different Skill Templates. Table 5 compares the evidence quality obtained using different skill templates for MCTS search. Human-based Skill, annotated by radiologists, consistently outperforms LLM-based Skill generated by GPT-5.5, indicating that expert prior knowledge provides more reliable guidance for evidence exploration than one-shot LLM generation. In comparison, CSC-SkillEvo achieves the best performance across all three metrics, further surpassing Human-based Skill. This advantage comes from combining counterfactual feedback with iterative self-evolution: instead of relying on a fixed skill template, CSC-SkillEvo continuously refines its skills using accumulated feedback and failure trajectories during search. Such dynamic adaptation enables the search policy to better align with evolving search states, leading to higher-quality evidence exploration.

Table 5: Comparison of evidence quality obtained using different skill templates on 100 cases from MIMIC-CXR and IU X-Ray."Human-based Skil" denotes a skill annotated by radiologists, while "LLM-based Skil" denotes a skill generated by GPT-5.5.

Skill Template	Evi-F1	CQ_{LLM}	CQ_{Rad}
Human skill	0.701	0.693	0.821
LLM skill	0.685	0.688	0.804
CSC-SkillEvo (Ours)	**0.723**	**0.719**	**0.846**

C.1.2 Robustness and Generalization Analysis

Recovery Performance after Early Planning Errors. To evaluate robustness to early planning errors in multi-tool evidence orchestration, we force each method to execute the worst action selected by its action-selection mechanism at the initial search stage. Table 6 reports the results. Beam Search achieves better recovery performance than DFSDT by maintaining multiple candidate trajectories, but this robustness comes with a higher Recovery Cost. In contrast, CSC-SkillEvo achieves the strongest recovery performance while maintaining the lower Recovery Cost. This advantage benefits from the counterfactual supervision accumulated during skill evolution. Consequently, CSC-SkillEvo recovers the optimal trajectory in 44% of cases within only four tool calls, and further increases the recovery rate to 78.5% by the end of the search, indicating that its recovery advantage persists as additional evidence is acquired.

Table 6: Recovery performance across exploration strategies after early planning errors on 100 cases from MIMIC-CXR and IU X-Ray. RR(n) denotes the proportion of cases that recover the optimal trajectory within n additional tool calls; RR(F) denotes the final recovery rate under the full remaining budget; Recovery Cost is the average number of additional tool calls required for successful recovery."*" denotes methods re-implemented from the original papers for clinical evidence extraction.

Method	RR(4)	RR(5)	RR(6)	RR(F)	Recovery Cost
Beam Search	15.0%	21.0%	23.0%	59.0%	6.252
DFSDT* (Qin et al., 2024)	7.0%	12.0%	13.0%	51.0%	4.548
MCTS	43.0%	53.0%	63.0%	70.0%	4.858
CSC-SkillEvo (Ours)	44.0%	65.0%	66.5%	78.5%	4.729

Robustness to Clinical Evidence Perturbations during MCTS. To assess the robustness of evidence exploration in our multi-tool evidence orchestration framework, we conduct the perturbation study reported in Table 7. Specifically, given the anchor findings identified by the Classifier tool, "Target" randomly selects one anchor finding and perturbs its corresponding clinical observation in the clinical evidence E. "Random" instead perturbs the clinical observation associated with a randomly selected non-anchor finding, while "None" applies no perturbation. Action Consistency measures the proportion of actions that remain unchanged after perturbation, whereas Trajectory Consistency measures the proportion of optimal search trajectories that are preserved. Compared with "Base+MCTS", "Base+CSC-SkillEvo" exhibits more stable evidence-search behavior under both the unperturbed and perturbed settings. The results under the two perturbation schemes further show that CSC-SkillEvo enables MCTS to adapt its tool calls and search trajectories in response to changes in the available evidence, thereby improving the robustness of MCTS exploration to potentially noisy clinical evidence and strengthening its ability to recover from erroneous intermediate search decisions.

Table 7: Robustness of MCTS evidence exploration via evidence perturbations on 100 cases from MIMIC-CXR and IU X-Ray.

Method	Finding Perturbation Setting	Action Consistency	Trajectory Consistency	Evi-F1	CQ_{LLM}	CQ_{Rad}
Base+MCTS	None	–	–	0.606	0.630	0.692
	Target	0.472	0.283	0.573	0.602	0.667
	Random	0.493	0.310	0.591	0.619	0.672
Base+CSC-SkillEvo	None	–	–	0.723	0.719	0.846
	Target	0.488	0.296	0.701	0.689	0.812
	Random	0.512	0.321	0.713	0.704	0.828

C.1.3 Reasoning Internalization Analysis

Performance of Different Reasoning Internalization Strategies. As shown in Table 8, we compare different strategies for internalizing clinical evidence reasoning under varying training and inference conditions. "Evidence Prompting," which introduces evidence only at inference time, already outperforms the "Base" model, confirming the value of explicit clinical evidence for report generation. "Evidence-Augmented SFT" further yields more consistent improvements in generation quality and clinical consistency than "GRPO" suggesting that different internalization mechanisms vary in effectiveness. In contrast, CA-OPSD achieves the best overall performance on BLEU, METEOR, ROUGE-L, GREEN, and Rad-F1 by leveraging the multi-source clinical evidence collected by MCTS during training to transfer and internalize diagnostic reasoning into the student policy. This enables more effective use of evidence supervision while reducing reliance on explicit evidence at inference time, resulting in better linguistic quality and clinical consistency.

Table 8: Comparison of different evidence internalization strategies.

Internalization Strategies	Training Condition	Inference Condition	BL-4	MTR	RG-L	GREEN	Rad-F1
Base	Image only	Image only	0.195	0.201	0.358	0.595	0.408
Evidence Prompting	Image only	Image+E_x^+	0.221	0.218	0.372	0.616	0.428
Evidence-Augmented SFT	Image+E_x^+	Image only	0.248	0.223	0.413	0.617	0.436
GRPO	Image+E_x^+	Image only	0.233	0.226	0.389	0.613	0.433
CA-OPSD(Ours)	**Image+$E_x^+ + E_x^-$**	**Image only**	**0.267**	**0.239**	**0.428**	**0.631**	**0.450**

Comparison of Negative Evidence Selection Strategies. As shown in Table 9, different negative evidence selection strategies exhibit a consistent trend: as the selection becomes more selective from Top-3 to Top-1, BL-4, MTR, RG-L, GREEN, and Rad-F1 all improve steadily. CA-OPSD with Top-1 negative evidence achieves the best performance across all metrics, with particularly pronounced gains on clinical metrics such as GREEN and Rad-F1. These results indicate that higher-ranked negative evidence provides more discriminative contrastive supervision, enabling the model to better distinguish positive from negative evidence during training while simultaneously improving both linguistic quality and clinical consistency.

Table 9: Comparison of hard-negative evidence selection strategies for CA-OPSD on IU X-Ray. Top-n denotes using the top-n ranked samples as hard-negative evidence for CA-OPSD.

Hard-negative Evidence	BL-4	MTR	RG-L	GREEN	Rad-F1
Top-3	0.225	0.232	0.394	0.613	0.421
Top-2	0.230	0.235	0.401	0.614	0.423
Top-1 (Ours)	**0.267**	**0.239**	**0.428**	**0.631**	**0.450**

Comparison with Different OPSD Methods. OPSD directly distills supervision from the reference report. PW-OPSD further introduces fixed position-dependent token weights to modulate the distillation strength, while DASH dynamically constructs token weights by combining sequence-relative divergence, adaptive propagation gates, and backward multi-step aggregation. As shown in Table 10, PW-OPSD and DASH yield consistent improvements on MIMIC-CXR and IU X-Ray through position-aware weighting and dynamic distillation, respectively. In comparison, CA-OPSD achieves the best performance across all evaluation metrics, demonstrating its stronger ability to align generated reports with clinical evidence, reduce factual errors, and improve both report accuracy and clinical reliability.

Table 10: Comparison of different OPSD methods on MIMIC-CXR and IU X-Ray."*" indicates that the original method is re-implemented within our framework.

	MIMIC-CXR						IU-X-Ray				
Methods	**BL-4**	**MTR**	**RG-L**	**P**	**R**	**F1**	**BL-4**	**MTR**	**RG-L**	**GREEN**	**Rad-F1**
OPSD* (Zhao et al., 2026a)	0.202	0.194	0.340	0.575	0.529	0.551	0.208	0.200	0.361	0.602	0.395
PW-OPSD* (Liu et al., 2026d)	0.206	0.198	0.345	0.582	0.534	0.557	0.234	0.204	0.396	0.623	0.424
DASH* (Hou et al., 2026)	0.210	0.203	0.347	0.589	0.537	0.562	0.242	0.207	0.398	0.618	0.432
CA-OPSD(Ours)	**0.224**	**0.223**	**0.355**	**0.604**	**0.554**	**0.572**	**0.267**	**0.239**	**0.428**	**0.631**	**0.450**

C.2 Additional Qualitative Results

Effect of Different Curriculum Stage Combinations. Figure 9 visualizes the impact of different curriculum-stage combinations in CA-OPSD, with darker colors indicating better performance. Single-stage training consistently underperforms the full model, suggesting that isolated knowledge transfer is insufficient to bridge the teacher–student gap. As more stages are introduced, performance steadily improves, showing that evidence discrimination, evidence grounding, and reasoning consolidation complement each other during progressive training. The full three-stage CA-OPSD achieves the best overall results, facilitating the student policy's internalization of critical clinical knowledge and improving report generation quality.

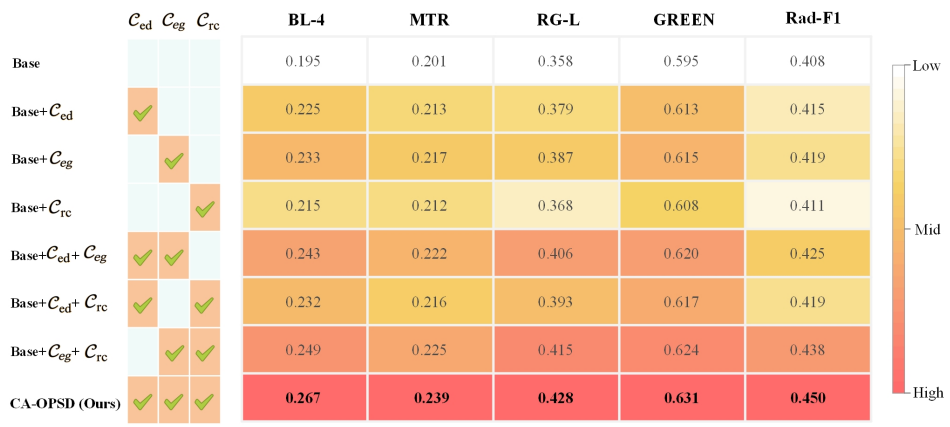

Figure 9: Performance of different curriculum-stage combinations in CA-OPSD on IU X-Ray. $\mathcal{C}_{ed}$, $\mathcal{C}_{eq}$, and $\mathcal{C}_{rc}$ denote the evidence discrimination, evidence grounding, and reasoning consolidation stages, respectively.

Evidence Quality vs. Search Cost. Figure 10 compares the Evidence F1 of DFSDT, ToT, MCTS, and CSC-SkillEvo under different simulation budgets. Evidence quality improves as the budget increases, but the gains diminish beyond about 30 simulations. CSC-SkillEvo rises from 0.470 at 5 simulations to 0.723 at 30, while increasing the budget to 40 yields only a marginal improvement to 0.742. At 30 simulations, CSC-SkillEvo also outperforms MCTS at 0.616, as well as ToT and DFSDT. These results show that most gains are achieved within 30 simulations, and that CSC-SkillEvo provides higher quality clinical evidence under the same search cost, yielding a better accuracy efficiency tradeoff.

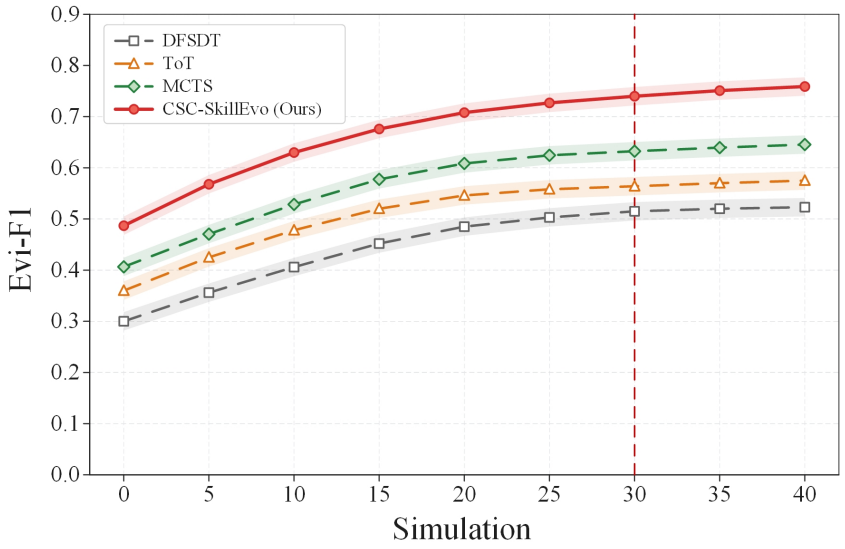

Figure 10: Comparison of evidence quality across different search methods under varying simulation budgets. Evidence F1 measures the quality of the retrieved clinical evidence. The dashed line marks 30 simulations, where the performance gains begin to plateau noticeably.

Visualization of Evidence Orchestration Performance during CSC-SkillEvo. As shown in Figure 11, Evi-F1improves stepwise from 0.436 to 0.748 during CSC-SkillEvo. Gains occur only in selected batches, suggesting that performance improves when newly identified failure patterns in tool prioritization and search termination are corrected, rather than simply through additional op-

timization rounds. The final plateau indicates that major search failure modes have largely been addressed and the evolved Skills are approaching convergence.

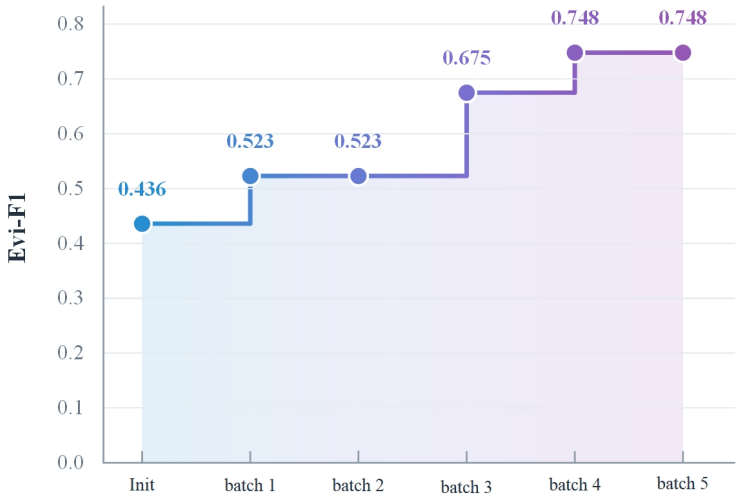

Figure 11: Evidence Quality during CSC-SkillEvo. Using 50 randomly sampled cases from IU X-Ray, we evaluate how batch-wise updates to the tool-prioritization and search-termination skills affect evidence quality, where "Init" denotes the initial skills before evolution.

Effect of Different VLM backbones for the Skill Optimizer. Figure 12 evaluates the sensitivity of CSC-SkillEvo to different VLM backbones. With all subsequent CSC-SkillEvo configurations fixed, every backbone successfully evolves the tool-prioritization and search-termination Skills, although the resulting Skill quality varies substantially. Open-source backbones generally yield lower scores across all three metrics, whereas closed-source backbones perform consistently better, suggesting that extracting actionable search experience from failure trajectories requires strong reasoning and abstraction capabilities. Among them, GPT-5.5 achieves the best performance, with 0.748 Evi-F1, 0.719 LLM-Judge, and 0.846 Physician score. Notably, these Skills are evolved on the CheXpert-plus dataset [1] and directly transferred to IU X-Ray for evaluation, indicating that CSC-SkillEvo is not limited to the source dataset but can support stable evidence exploration and high-quality evidence acquisition across datasets.



Figure 12: Comparison of Different VLM backbones for the Skill Optimizer in CSC-SkillEvo on 50 Cases Sampled from IU X-Ray.

Visualization of Tool-Prioritization and Search-Termination Skill Evolution. To qualitatively analyze the effectiveness of skill self-evolution, Figure 13 illustrates how CSC-SkillEvo progressively refines the tool-prioritization and search-termination skills through counterfactual feedback. At the initial stage, tool-prioritization ranks candidate tools using fixed scoring heuristics based on factors such as tool relevance and expected evidence gain, while search-termination determines whether to stop the search according to evidence coverage and predefined stopping criteria. Such static rules are inherently limited in their ability to adapt to the evolving state of a dynamic search process. After the evolution to version 2, the skills begin to incorporate search feedback, enabling tool selection to adapt to the current evidence state and termination decisions to jointly account for unresolved findings and potential information gain. In the final version, the applicability conditions of these rules are further refined: tool calls are explicitly directed toward complementary verification of missing evidence, while the search is terminated promptly once critical evidence is sufficiently established and the expected benefit of further exploration becomes marginal. Overall, this progression demonstrates how CSC-SkillEvo evolves from static heuristic rules toward an experience-driven capability for adaptive search control.

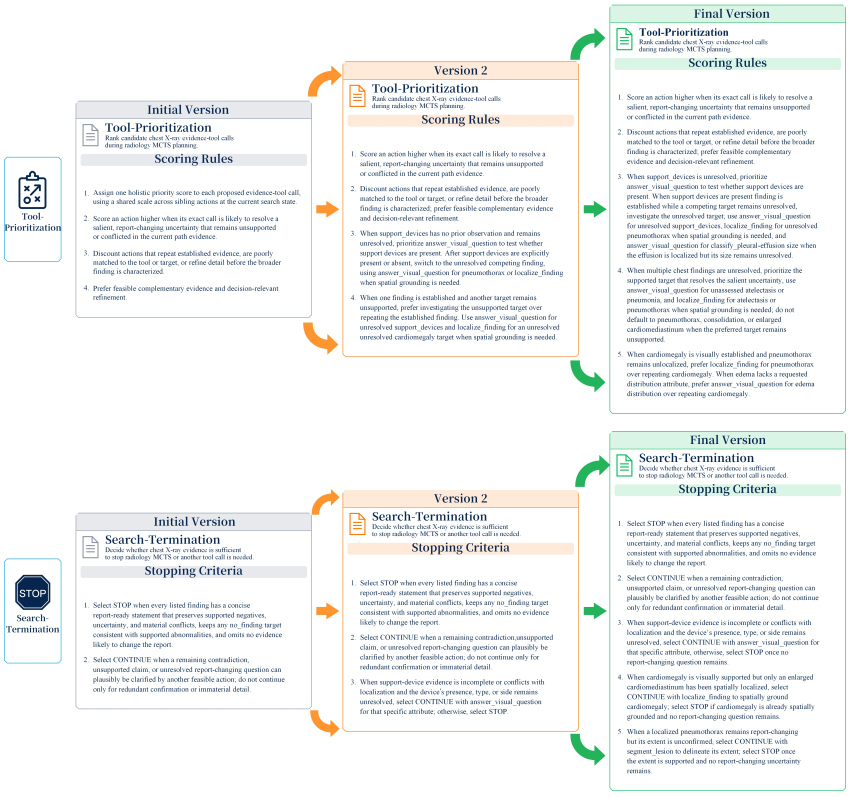

Figure 13: Iterative Evolution of Tool-Prioritization and Search-Termination Skill Prompts in CSC-SkillEvo.

C.3 Case Studies

Qualitative Comparison of Positive and Negative Evidence. Figure 14 compares the positive and negative evidence generated by CSC-SkillEvo against the ground-truth. Positive evidence accurately captures the key left-lung opacity, further localizes it through grounding, and preserves other clinically relevant findings, yielding strong agreement with the ground-truth. In contrast, negative evidence contains superficially relevant descriptions but misses the primary abnormality and introduces weakly supported focal findings, resulting in lower clinical consistency. This comparison shows that seemingly relevant search trajectories can differ substantially in diagnostic value, while CSC-SkillEvo effectively favors evidence better aligned with the underlying diagnosis. It further supports using positive and hard-negative evidence for contrastive supervision, improving the

model's ability to distinguish reliable clinical evidence from misleading information and providing a stronger basis for subsequent evidence-grounded reasoning internalization.

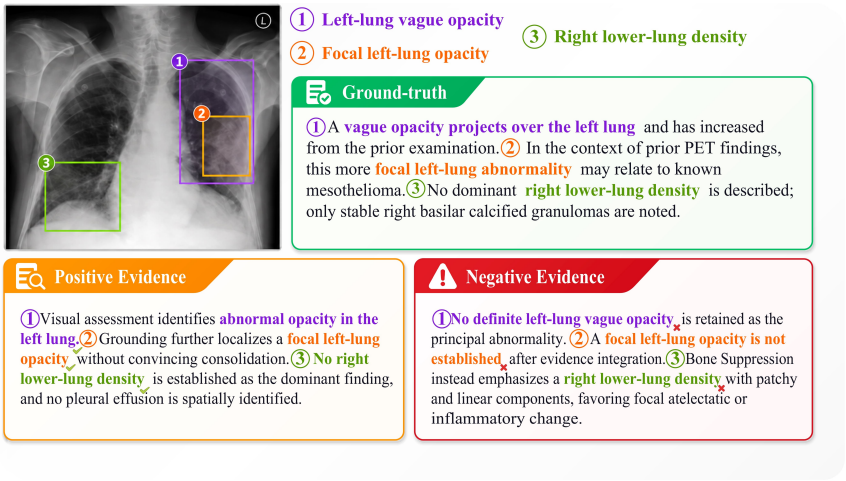

Figure 14: Qualitative comparison of positive and negative evidence in a representative MIMIC-CXR case. Purple, orange, and green regions denote left-lung hazy opacity, focal left-lung opacity, and increased right lower-lung density, respectively. Red crosses and green check marks indicate incorrect and correct evidence.

Qualitative Comparison of Generated Reports Across Ablation Modules. Figure 15 compares the ground-truth report with those generated by different ablation variants on the IU X-Ray dataset. The "Base" model captures salient findings such as cardiomegaly, but misses atelectasis and introduces unsupported consolidation-related descriptions. Adding "Evidence Orchestration (EO)" improves the generation of negative findings, such as the absence of pneumothorax, but still fails to capture atelectasis, suggesting limited support for fine-grained positive abnormalities. In contrast, our full model accurately identifies both cardiomegaly and mild bibasilar atelectasis, while also generating more complete negative findings, including no focal airspace consolidation, pleural effusion, or pneumothorax. Overall, our method produces reports that are more consistent with the ground truth in both clinical content and semantic structure.

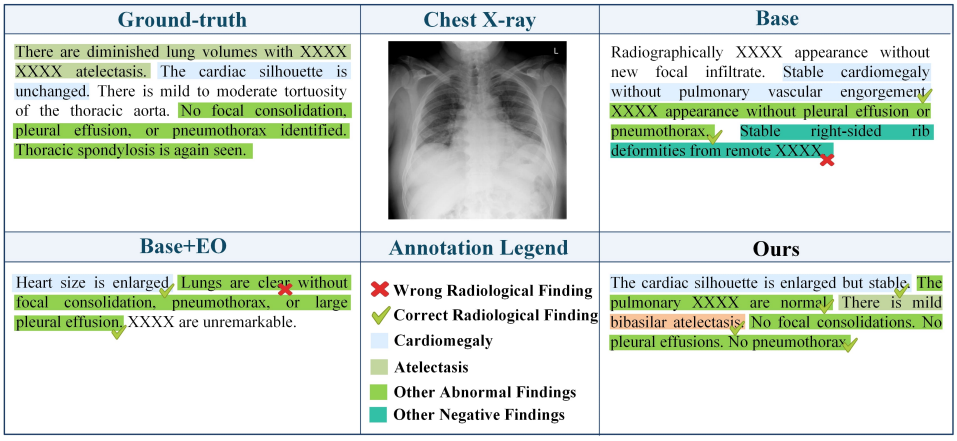

Figure 15: Qualitative comparison of generated reports on IU X-Ray. Colored highlights indicate clinical findings matched between the ground-truth report and the reports generated by different models. Red crosses and green check marks denote incorrect and correct report descriptions, respectively.

Visualizations of Cross-Attention. Figure 16 compares the cross-attention distributions of CSC-SkillEvo and competing methods when generating descriptions of cardiomegaly and edema, together with the corresponding ground-truth lesion regions. Although the "Base" model identifies cardiomegaly, its attention remains diffuse and insufficiently covers the pulmonary regions associated with edema, leading to the omission of this finding. With "evidence prompting", the model successfully generates edema-related descriptions, yet its attention still fails to consistently localize the regions associated with cardiomegaly. This suggests that the improvement largely stems from external evidence injection rather than enhanced intrinsic visual grounding and evidence reasoning. In contrast, CSC-SkillEvo captures all abnormalities in the GT while consistently attending to diagnostically relevant regions. These results demonstrate that Skill-guided sequential evidence search and orchestration improve the utilization of visual evidence, reduce missed clinical findings, and strengthen cross-modal consistency between generated reports and their supporting image evidence.

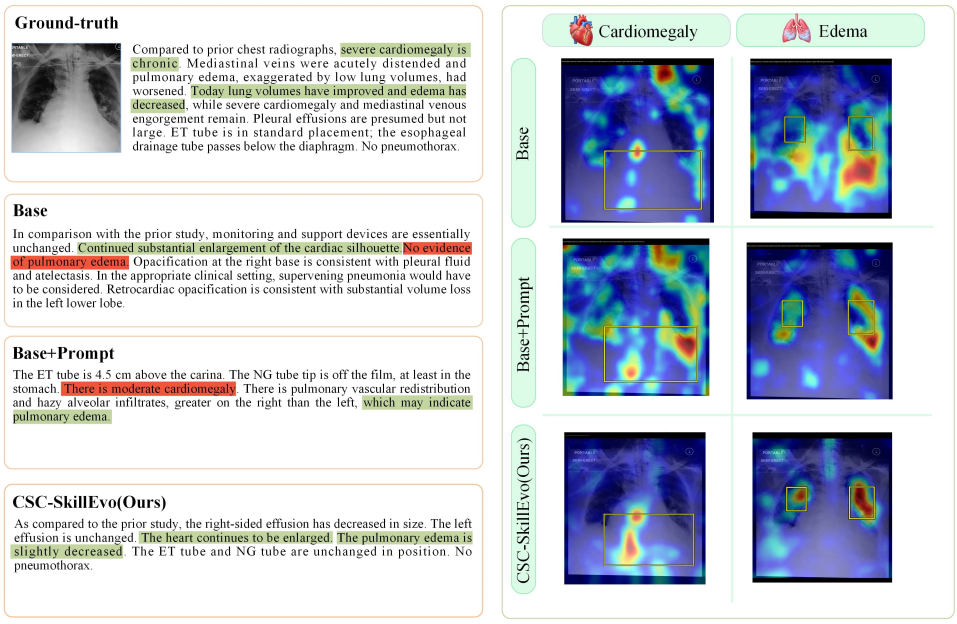

Figure 16: Qualitative comparison of generated reports and cross-attention maps on MIMIC-CXR. Left: ground-truth and generated reports, with green/red indicating consistent/conflicting descriptions. Right: cross-attention maps for clinical finding tokens, overlaid with ground-truth localization regions.

Visualization of Clinical Evidences. Figure 17 compares the clinical evidence produced by different evidence search strategies, showing that effective search leads to more comprehensive imaging-evidence coverage and better capture of key findings. The Greedy strategy identifies abnormalities related to lung opacity, but its locally optimal search behavior limits further exploration of other findings and exclusionary evidence. ReAct acquires richer evidence through interactive reasoning and provides explanations related to lung opacity and pneumonia, yet its evidence remains relatively narrow and lacks broader support. DFSDT further extends evidence acquisition through multi-step search and incorporates Grad-CAM for spatial interpretation; however, inconsistencies between its activation regions and the annotated masks lead to pneumonia inferences supported by insufficient visual evidence. In contrast, our method achieves more thorough evidence search and orchestration, accurately identifies the abnormal region in the right lower lung, and integrates multi-source observations to rule out unsupported conclusions, resulting in more consistent vision–text alignment. These results indicate that the proposed method yields more reliable imaging-evidence representations and improves the clinical consistency of the Evidence Report.

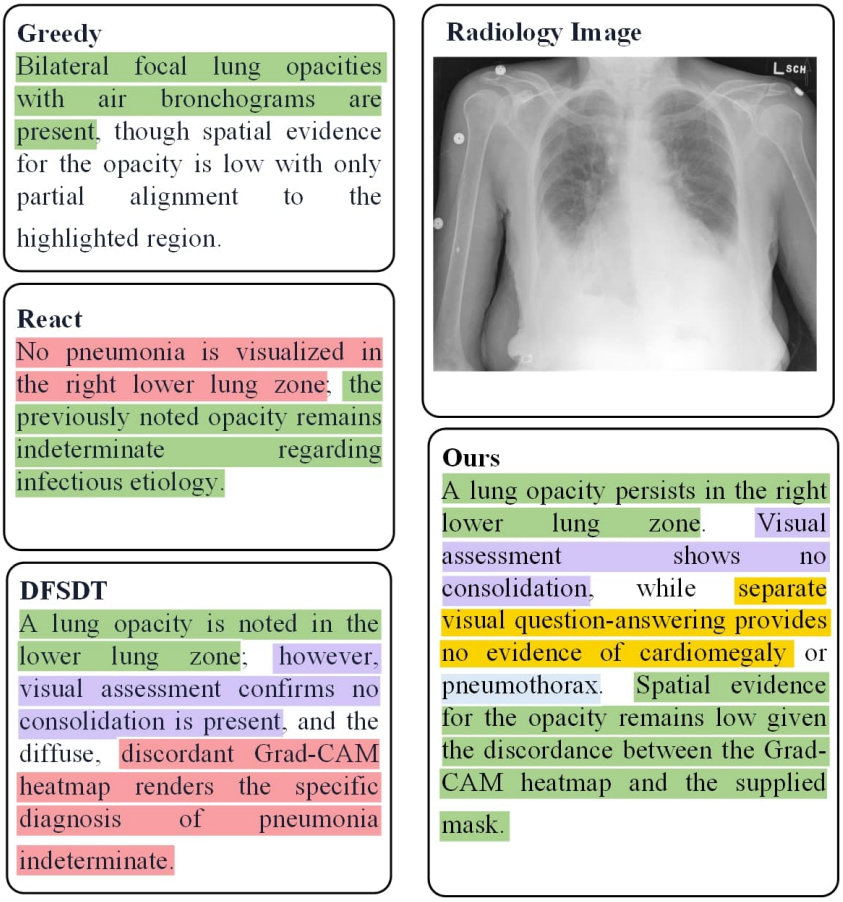

Figure 17: Qualitative visualization of clinical evidences obtained by different search strategies. The figure compares the clinical evidence generated by four search strategies: Ours, ReAct, DFSDT, and Greedy. Different highlight colors indicate evidence snippets that are relevant to the corresponding finding descriptions.

D Limitations and Discussion

Although the current radiology tool library $\mathbb{T}_{\text{rad}}$ covers complementary functions for global diagnosis, localized analysis, and finding-level verification, it integrates only a limited number of off-the-shelf radiology tools. The limited coverage of the current tool space may constrain the ability of MCTS to explore diverse clinical evidence, particularly for lesions and abnormal findings that require more specialized analysis. Future work could incorporate a broader range of radiology models, external retrieval resources, and task-specific specialized tools to expand the candidate evidence space and further strengthen multi-tool evidence orchestration. Notably, even under the current relatively compact tool library setting, our method achieves state-of-the-art performance on the evaluated RRG benchmarks. Meanwhile, both objective and subjective evaluations of clinical evidence show that our evidence orchestration method can acquire clinical evidence with higher clinical quality and factual accuracy. Taken together, these results suggest that improvements in RRG performance do not solely depend on increasing the number of external tools, but also benefit from effective multi-tool evidence orchestration and subsequent reasoning internalization. Therefore, constructing a larger and more diverse radiology tool ecosystem may further broaden the accessible clinical evidence space and improve the capability ceiling and generalization potential of the proposed framework.