

MedSPA: Self-planning Medical Agents

Luu Van Huy, Bowen Wang*, Fumio Okura, and Yuta Nakashima

The University of Osaka, Japan

{luu, wang, okura, n-yuta}@im.sanken.osaka-u.ac.jp

Abstract

Reliable medical diagnosis requires a precise final prediction, which is strictly supported by observable evidence. Recent large vision-language models perform well on medical image understanding tasks, yet they often lack explicit control over which evidence to inspect before making a prediction. To facilitate flexible planning of evidence exploration, we propose MedSPA, a self-planning medical reasoning framework that separates diagnostic evidence acquisition from evidence-grounded prediction with two cooperative agents: a planning-and-reasoning agent that adaptively decides what to inspect and when to stop, and a summary agent that generates the final output from the accumulated evidence. MedSPA distills medical knowledge from strict diagnostic procedures but acquires flexibility through reinforcement learning to generate optimal outputs. Experiments on medical report generation and medical visual question answering show that MedSPA improves clinical performance, faithfulness, and controllability over existing reasoning baselines.

1 Introduction

Recent medical vision-language models (Zhang et al., 2024) have achieved impressive performance across various medical tasks (Li et al., 2023; Wang et al., 2024; Huppertz et al., 2025). Beyond final prediction accuracy, reliable medical diagnosis also requires models to expose how diagnostic conclusions are supported by visual evidence. This has motivated growing interest in medical vision-language reasoning, where models are expected not only to produce an answer or report, but also to organize intermediate evidence in a clinically meaningful way (Xia et al., 2024; Fansi Tchango et al., 2022).

Chain-of-thought prompting (Wei et al., 2022) and its variants (Kojima et al., 2022) advance this

*Corresponding author.

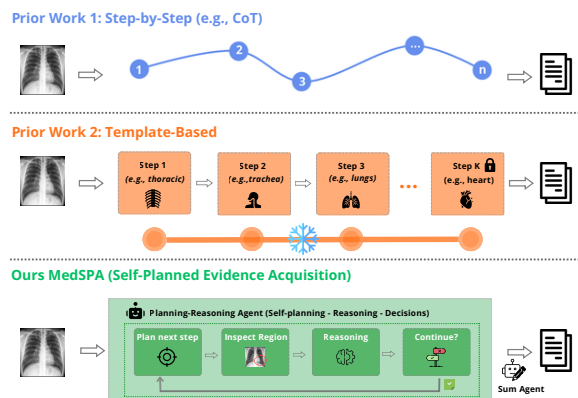


Figure 1: Conceptual comparison of diagnostic reasoning paradigms. **CoT-style reasoning** offers flexible step-wise exploration, whereas **Template-based reasoning** provides clinical structure through predefined diagnostic procedures. **MedSPA** integrates these complementary strengths by converting rigid clinical templates into adaptive self-planned evidence acquisition, enabling reasoning that is both flexible and clinically grounded.

goal by encouraging models to decompose complex tasks into intermediate reasoning steps, thereby making the prediction process more transparent and helping models handle complex diagnostic requirements (Figure 1(top)). In medical imaging, such reasoning chains can become more reliable when they are guided by clinical reasoning patterns from real diagnostic practice. Medical templates offer this guidance by injecting domain knowledge, such as expert-designed search strategies, anatomical checklists, and diagnostic criteria, into models, forcing them to inspect images in a more systematic manner (Singh et al., 2015; Herring, 2019) as in Figure 1(middle).

In real-world diagnostic practice, however, human clinicians do not simply traverse such templates as fixed checklists. They use such standardized diagnostic routines as clinical priors while dynamically adjusting their diagnostic focus based on their knowledge and the evidence observed so far.

This motivates a framework that preserves the clinical structure of medical templates while enabling case-adaptive diagnostic planning.

A key question arises: *How can a model balance between the explicit knowledge in the templates and intrinsic but general knowledge in the model, like human clinicians do?*

To answer this question, we propose MedSPA, a self-planned medical reasoning framework that distills knowledge in medical templates into the model and learns to adaptively plan the reasoning process through reinforcement learning. MedSPA consists of two agents (Figure 1(bottom)): the planning-and-reasoning agent decides what evidence to inspect in the given image and when to stop the reasoning process, while the summary agent generates the final report or answer from the accumulated evidence from the planning-and-reasoning agent.

Contribution. Our experimental results demonstrated that MedSPA consistently achieved strong clinical performance among existing methods, and diagnostic plans generated by MedSPA efficiently extract sufficient medical evidence from the given images. These results suggest that knowledge distillation from medical templates and learning to ignore them through reinforcement learning offer a good balance between rigorous medical standards and uncontrolled chain-of-thought. This new insight can serve as a strong guideline for designing agentic systems for knowledgeable domains.

2 Related Work

2.1 Medical Vision-Language Reasoning

Large-scale medical VLMs (Li et al., 2023; Zhang et al., 2023) and general-purpose multimodal large language models (MLLMs) (Wang et al., 2026; Saab et al., 2024) have recently demonstrated increasing potential for medical image understanding, including radiology report generation (Bannur et al., 2024) and medical visual question answering (Eslami et al., 2023; Singhal et al., 2023). Despite their advancements, existing medical VLMs still face difficulty in producing reliable diagnostic reasoning and clinically faithful explanations (Zhu et al., 2025; Moll et al., 2026). To alleviate these issues, recent studies have explored stepwise reasoning strategies (Wei et al., 2022; Wang et al., 2023a) that have shown promising success in general reasoning tasks, aiming to encourage more plausible rationales when describing medical images (Liu

et al., 2024b). However, even with such strategies, the lack of supervision that teaches models to reason over medical images like real clinicians remains a major obstacle to clinically grounded reasoning. In this work, we curate a dataset with explicit planning and reasoning steps, aiming to train models to acquire planning-oriented diagnostic reasoning skills and adapt to diverse medical cases.

2.2 Medical Reasoning Processes Optimization

Reinforcement learning (Mnih et al., 2015) and reward modeling (Rafailov et al., 2023) have recently been used to improve the reasoning behavior of medical large vision-language models. However, medical datasets usually provide only final reports or answers, while supervision for intermediate diagnostic evidence acquisition is rarely available. When a final prediction is incorrect, it is often unclear whether the error is due to missing evidence, flawed reasoning, or a final summary not grounded in the evidence. Recent works attempt to address this delayed rewards problem through process-level feedback (Zhu et al., 2024) or reward assignment to generated reasoning trajectories during post-training (Lai et al., 2026; Yun et al., 2025). MedSPA follows this reward-based strategy, but differs in its role-specific reward design, which enables each agent to learn behaviors aligned with its own responsibility rather than being optimized only for the final outcome.

3 Methodology

MedSPA consists of two agents based on a large vision-language model (LVLM) with distinct roles: A Planning and Reasoning (PR) agent and a Summary (SUM) agent. These two agents are trained to balance between following strict diagnostic standards represented by templates and free-form CoT reasoning.

We instantiate this balance via two-stage training: *medical knowledge distillation* and *role-specific reinforcement alignment*. For medical knowledge distillation (§3.2), we introduce an automatic knowledge curation pipeline that converts image-annotation pairs into structured reasoning chains, as shown in Figure 2(top), where the annotations are a diagnostic report for medical report generation (MRG) or an answer for medical visual question answering (MedVQA). These cu-

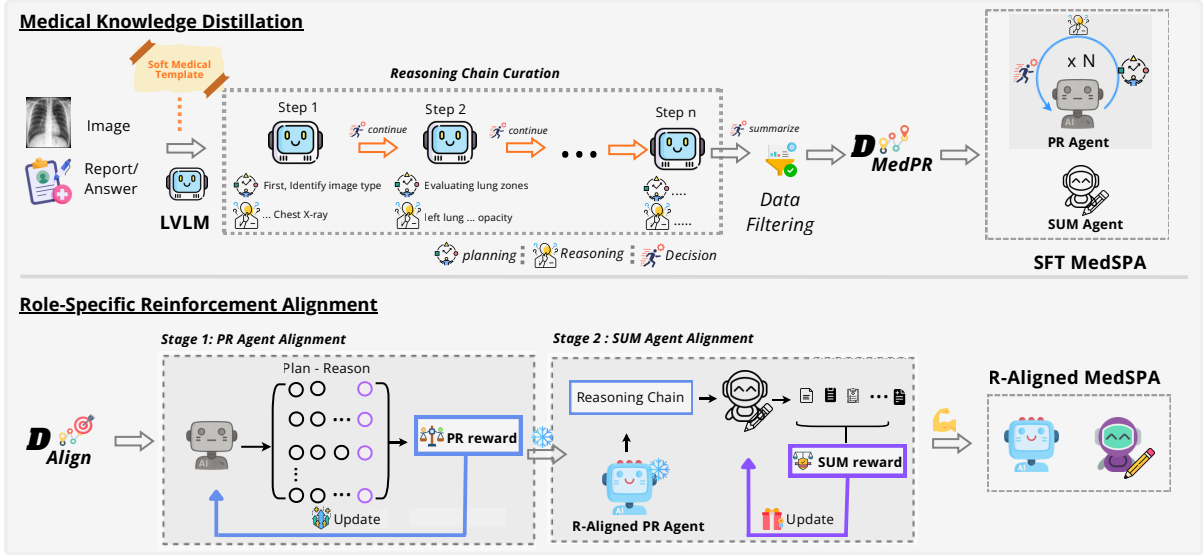


Figure 2: Training pipeline of the proposed MedSPA framework.

rated chains are used for supervised fine-tuning of both agents. For role-specific reinforcement alignment (§3.4), we apply a reinforcement learning framework that aligns evidence acquisition and final prediction with their corresponding objectives, as illustrated in the bottom panel of Figure 2.

3.1 Inference

Given an input medical image I , MedSPA sequentially runs the PR and SUM agents, where the former is formulated as an iterative evidence-acquisition loop to explore I for medical findings step-by-step, and the latter summarizes the evidence from the PR agent to generate the final output for I .

PR Agent. In each iteration of the evidence-acquisition loop, the PR agent first determines a diagnostic focus (i.e., *action*) and looks into I for some medical evidence to make medical reasoning (i.e., *reasoning*). It then decides whether to acquire additional evidence or not (i.e., *decision*).¹

Formally, at iteration t , the PR agent produces an action-reasoning-decision triple (a_t, r_t, d_t) , where a_t denotes the selected diagnostic focus, such as examining a specific anatomical region or verifying a suspected abnormality; r_t represents the image-grounded reasoning associated with the selected focus; and $d_t \in \{\text{continue}, \text{summarize}\}$ indicates whether the agent should continue the evidence-acquisition loop or terminate it and proceed to summary generation.

The PR agent also maintains a reasoning memory $M_t = \{(a_k, r_k, d_k)\}_{k=1}^{t-1}$, where $M_1 = \emptyset$ at the first iteration ($t = 1$). This memory is used as the context for subsequent iterations, recording previous reasoning steps and decisions to help the PR Agent avoid redundant inspection and progressively cover the diagnostic evidence needed for final output generation. After each iteration, the newly generated triple (a_t, r_t, d_t) is appended to the memory.

Once the agent decides to terminate the loop at $t = T$ ($d_T = \text{summarize}$), the memory M_{T+1} is fed to the SUM agent as the diagnostic reasoning chain is defined as $Z = \{(a_t, r_t, d_t)\}_{t=1}^T$.

SUM Agent. The SUM agent takes I and Z as input to generate the task-specific output $\hat{o}$, where $\hat{o}$ can be medical report for MRG or predicted answer for MedVQA.²

3.2 Medical Knowledge Distillation

To inject the standard knowledge into our PR and SUM agents, we first curate medical reasoning chains using a strong LVLM instructed to follow diagnostic templates to generate outputs. The curated dataset is used to fine-tune the agents in a supervised manner.

Reasoning Chain Curation Let $\mathcal{D} = \{(I, y)\}$ denote medical image-annotation pairs, where y is a set of report sentences for MRG and is a question-answer pair (q, u) for MedVQA. We leverage an LVLM to generate a reasoning chain

¹The prompt for the PR agent is in Appendix A.1.

²The prompt can be found in Appendix A.2.

$Z = \{(a_t, r_t, d_t)\}_{t=1}^T$ for each pair $(I, y) \in \mathcal{D}$. To guide chain generation, we design a medical reasoning prompt p^3 that incorporates a soft medical reasoning template developed from radiology textbooks and consultations with board-certified clinicians. The template encourages a global assessment of I followed by focused examination of suspicious regions or findings, while prompting explicit continue-or-summarize decisions, thereby providing expert-informed guidance.

Let $p(y)$ denote the prompt to generate Z , where the prompt is augmented with the ground-truth annotation y , which serves as a semantic hint toward clinically relevant findings. We obtain the chain as $Z = \text{LVL}(I, p(y))$ for all $(I, y) \in \mathcal{D}$ to construct a curated dataset, denoted as $\mathcal{D}_{\text{MedPR}} = \{(I, y, Z)\}$.

Quality Control. Because automatically generated reasoning chains may contain erroneous reasoning, like ungrounded or incomplete findings, we apply a two-stage filtering process to ensure the reliability of generated chains.

Following (Irvin et al., 2019), we define functions $\phi(Z) \in \{0, 1\}^{|\mathcal{L}|}$ and $\psi(y) \in \{0, 1\}^{|\mathcal{L}|}$ to extract medically meaningful labels in the predefined label set $\mathcal{L}$, represented as multi-hot vectors. For all $(I, y, Z) \in \mathcal{D}_{\text{MedPR}}$, we compare $\phi(Z)$ and $\psi(y)$ to evaluate their alignment. We also employ an LLM-based evaluator to score Z based on its faithfulness to y and completeness of key clinical findings. Only chains that satisfy label consistency and exceed a predefined reasoning-quality threshold are retained in $\mathcal{D}_{\text{MedPR}}$. These checks provide semantic and clinical safeguards against unsupported intermediate reasoning, including hallucinated findings, incorrect anatomical locations, and missing clinically salient evidence. Some examples after quality control are found in the appendix.

3.3 Supervised Fine-tuning

We train the agents on $\mathcal{D}_{\text{MedPR}}$ via supervised fine-tuning, enabling the agents to capture clinically grounded planning/reasoning patterns from curated diagnostic chains.

PR Agent Training. Given $(I, y, Z) \in \mathcal{D}_{\text{MedPR}}$ and reasoning step t , we define a reasoning memory as $M_t = \{(a_k, r_k, d_k)\}_{k=1}^{t-1}$, which is a subset of the chain Z until step $t - 1$. Conditioned on I and M_t , the PR agent is trained to generate the triplet

$(a_t, r_t, d_t) \in Z$ at step t via autoregressive generation, rather than producing the complete reasoning chain in a single pass.

SUM Agent Training. The SUM agent is trained to generate a task-specific output. For MRG, it generates the reference diagnostic report y based on the corresponding I and Z in $\mathcal{D}_{\text{MedPR}}$; for Med-VQA, it produces the ground-truth answer u from I , Z and q .

3.4 Role-Specific Reinforcement Alignment

Although supervised fine-tuning forces agents to mimic $Z \in \mathcal{D}_{\text{MedPR}}$, it does not directly optimize for efficiency, i.e., whether the already acquired evidence is sufficient for accurate outputs, and for effectiveness to the final output, i.e., whether the evidence by the PR agent is retained or discarded for a better final output. We therefore introduce role-specific reinforcement alignment (R-Align), which assigns dedicated objectives to the PR and SUM agents.

We construct a dedicated dataset $\mathcal{D}_{\text{R-Align}} = \{(I, \psi(y))\}$ from $(I, y) \in \mathcal{D}$ for R-Align. Each agent learns to adjust its behavior from a role-specific reward computed from $\psi(y)$.

3.4.1 PR Agent Optimization

Given a sample $(I, \psi(y)) \in \mathcal{D}_{\text{R-Align}}$, the PR agent is used to sample a set $\mathcal{Z} = \{Z\}$ of $|\mathcal{Z}|$ different reasoning chains. For each chain $Z \in \mathcal{Z}$, we optimize the PR agent using rewards computed from the labels in $\phi(Z)$, rather than $\psi(\hat{o})$ from the task-specific output $\hat{o}$ by the SUM agent based on Z . This ensures that errors introduced by the SUM agent are not attributed to the PR agent’s evidence acquisition.

The reward for the PR agent is a combination of diagnostic consistency, structural validity, and coverage penalty. Specifically, the diagnostic consistency measures consistency between the labels in $\phi(Z)$ and $\psi(y)$:

$$R_{\text{con}}(Z, y) = \frac{\phi(Z)^\top \psi(y)}{|\mathcal{L}|}. \quad (1)$$

To preserve the format of each reasoning step produced by the PR agent, we use a structural validity reward:

$$R_{\text{fmt}}(Z) = \delta(Z), \quad (2)$$

where $\delta(Z)$ gives 1 when all reasoning steps in Z follow a specific format to represent triplet (a_t, r_t, d_t) , and otherwise 0.

³The full prompt is provided in the Appendix B.

To prevent the PR agent from stopping before collecting sufficient evidence, we penalize premature termination according to the number of labels that remain uncovered, i.e.,

$$R_{\text{cov}}(Z, y) = -\frac{C(\phi(Z), \psi(y))}{|\psi(y)|} \quad (3)$$

where $C(\phi(Z), \psi(y))$ counts the number labels that are positive in $\psi(y)$ but are negative in $\phi(Z)$, and $|\psi(y)|$ is the number of positive labels in $\psi(y)$. $R_{\text{cov}}(Z) = 0$ when $|\psi(y)| = 0$.

The overall reward for the PR agent is:

$$R^R(Z) = R_{\text{con}}(Z) + R_{\text{fmt}}(Z) + \beta R_{\text{cov}}(Z), \quad (4)$$

where β controls the strength of the coverage penalty during evidence acquisition.

3.4.2 SUM Agent Optimization.

For each sample $(I, \psi(y)) \in \mathcal{D}_{\text{R-Align}}$, the frozen PR agent first produces a reasoning chain Z and label vector $\phi(Z)$. Conditioned on I and Z , the SUM agent samples a set $\mathcal{O} = \{\hat{o}\}$ of $|\mathcal{O}|$ different outputs.

The reward for the SUM agent is designed to encourage the final output to faithfully preserve the diagnostic evidence acquired by the PR agent, without making the SUM agent responsible for errors in the PR agent’s evidence exploration, which is given by

$$R^S(Z, y, \hat{o}) = \sum_{l \in \mathcal{L}} r_l(\phi(Z), \psi(y), \psi(\hat{o})), \quad (5)$$

where $r_l(\cdot, \cdot, \cdot)$ computes a reward signal for $l \in \mathcal{L}$ for all possible combinations of labels, which is defined in the Appendix.

Both stages use group-relative advantages and are optimized with GRPO and Kullback-Leibler regularization. The loss functions are provided in the Appendix.

4 Experiments

4.1 Experimental Setup

Datasets and Tasks. We evaluate MedSPA on two medical vision-language tasks: medical report generation (MRG) and medical visual question answering (MedVQA). Across both tasks, Qwen3-VL-32B (Bai et al., 2025) is used as the teacher model⁴ to generate reasoning chains. In MRG,

⁴The rationale for selecting Qwen3-VL-32B is provided in the Appendix B.2.

we use MIMIC-CXR (Johnson et al., 2019) to construct **MedPR-CXR** corresponding to $\mathcal{D}_{\text{MedPR}}$ from image-report pairs, and evaluate on the official MIMIC-CXR test split. To assess out-of-distribution generalization, we further evaluate the MIMIC-trained model on IU X-ray (Demner-Fushman et al., 2016), without using IU X-ray samples for training or curation. For MedVQA, we construct **MedPR-VQA** from image-question-answer triplets in VQA-RAD (Lau et al., 2018) and SLAKE (Liu et al., 2021). We evaluate representative organ- and modality-specific subsets, including chest X-ray, brain MRI, and abdominal CT, covering both closed-ended (C) and open-ended (O) questions.⁵

Implementation Details. We train the PR and SUM agents for each of the MRG and VQA tasks. They are initialized with our backbone, Qwen3-VL-8B-Instruct. All experiments are conducted on eight A100 GPUs with 80GB VRAM. We train both agents using LoRA with rank 64 (Hu et al., 2022) under bfloat16 mixed precision. For supervised fine-tuning, each agent is trained for 10 epochs on MRG. They are trained for 3 epochs on MedVQA. The SFT agents are denoted as **SFT MedSPA**, and the final model after R-Align is denoted as **R-Aligned MedSPA**. During the R-Align training, we sample four reasoning chains for the PR agent and four candidate outputs for the SUM agent per input. For comparison, we also trained our backbone with supervised fine-tuning (Direct SFT) so that it can generate a gold report/answer $y \in \mathcal{D}$ via CoT. Other hyperparameters are detailed in Appendix D.

Metrics. In MRG, we report clinical efficacy (CE) metrics, that is, precision, recall, and F1, based on clinical labels extracted from generated and reference reports (Irvin et al., 2019), i.e., $\psi(\hat{o})$ and $\psi(y)$, together with standard NLG metrics: BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), and METEOR (Banerjee and Lavie, 2005). For MedVQA, F1 is used for open-ended questions, while accuracy is used for closed-ended questions.

4.2 Main Results

Table 1 summarizes the scores on MRG across MIMIC-CXR and IU X-ray. On MIMIC-CXR, SFT MedSPA improves the Direct SFT baseline

⁵Closed-ended questions have predefined answer options; open-ended questions require free-form answers.

Table 1: Comparison of MedSPA and state-of-the-art methods on MIMIC-CXR and IU-Xray for medical report generation. Missing values are denoted by “-”. **B**, **R**, and **M** mean BLEU, ROUGE, and METEOR, respectively.

Dataset	Model	CE Metrics			NLG Metrics			
		Prec.	Rec.	F1	B-1	B-4	R-L	M
MIMIC-CXR	R2Gen (Chen et al., 2020)	0.333	0.273	0.276	0.353	0.103	0.277	0.142
	R2GenCMN (Chen et al., 2021)	0.334	0.275	0.278	0.353	0.106	0.278	0.142
	METransformer (Wang et al., 2023b)	0.364	0.309	0.311	0.386	0.124	0.291	0.152
	Med-LLM (Liu et al., 2024c)	0.412	0.373	0.395	-	0.128	0.289	0.161
	MA (Shen et al., 2024)	0.411	0.398	0.389	0.396	0.115	0.274	0.151
	RGRG (Tanida et al., 2023)	0.461	0.475	0.447	0.373	0.126	0.264	0.168
	I3 + C2FD (Liu et al., 2024a)	0.465	0.482	0.473	0.402	0.128	0.291	0.175
	PromptMRG (Jin et al., 2024)	0.501	0.509	0.476	0.398	0.112	0.268	0.157
	DART (Park et al., 2025)	0.520	0.546	0.533	0.437	0.137	0.310	0.175
	Direct SFT	0.483	0.466	0.474	0.298	0.090	0.231	0.132
	SFT MedSPA	0.549	0.543	0.546	0.334	0.125	0.247	0.152
	R-Aligned MedSPA	0.569	0.552	0.560	0.341	0.129	0.260	0.169
IU-Xray	R2Gen (Chen et al., 2020)	0.141	0.136	0.138	0.325	0.059	0.253	0.131
	R2GenCMN (Chen et al., 2021)	0.183	0.187	0.180	0.266	0.063	0.180	0.146
	PromptMRG (Jin et al., 2024)	0.213	0.229	0.211	0.401	0.098	0.281	0.160
	Direct SFT	0.182	0.221	0.200	0.215	0.060	0.176	0.132
	SFT MedSPA	0.322	0.252	0.283	0.311	0.084	0.253	0.147
	R-Aligned MedSPA	0.324	0.301	0.312	0.337	0.086	0.267	0.152

by 0.072 F1, and R-Aligned MedSPA further increases the gain to 0.086. A similar trend is observed on IU X-ray, where SFT MedSPA improves F1 by 0.083 over Direct SFT, and R-Aligned MedSPA reaches a 0.112 gain. Among prior methods that report CE metrics on IU X-ray, R-Aligned MedSPA achieves the best clinical performance. The gains in CE metrics suggest that self-planned evidence acquisition improves the preservation of clinically relevant findings and supports generalization beyond the training distribution. We additionally compare MedSPA with several strong general-purpose MLLMs on the MIMIC-CXR test set, including DeepSeek-VL2 (Wu et al., 2024), Llama-4-Scout-17B-16E (Meta AI, 2025), GPT-4V (OpenAI, 2023), and Gemini 2.5 Pro (Gemini Team, 2025). Detailed results are provided in Appendix F.1.

Figure 3 provides a qualitative example. Direct SFT produces a fluent report but introduces un-

ported findings and misses important evidence. In contrast, R-Aligned MedSPA dynamically decomposes the case into multiple diagnostic steps, accumulates image-grounded findings in the reasoning memory, and uses the SUM agent to generate a final report grounded in the collected evidence. This example illustrates that MedSPA improves diagnostic reasoning not by simply extending the reasoning chain, but by guiding the model to aggregate relevant evidence before generating an output.

For MedVQA, we compared only the SFT version of MedSPA with some baseline methods. Since most questions in the evaluated MedVQA benchmarks require short answers such as yes/no responses or brief phrases, the current R-Align rewards⁶ are not directly applicable to MedVQA. Table 2 shows that MedSPA consistently improves the

⁶R-Align is designed for report generation, where its role-specific rewards are defined around evidence coverage and diagnostic consistency.

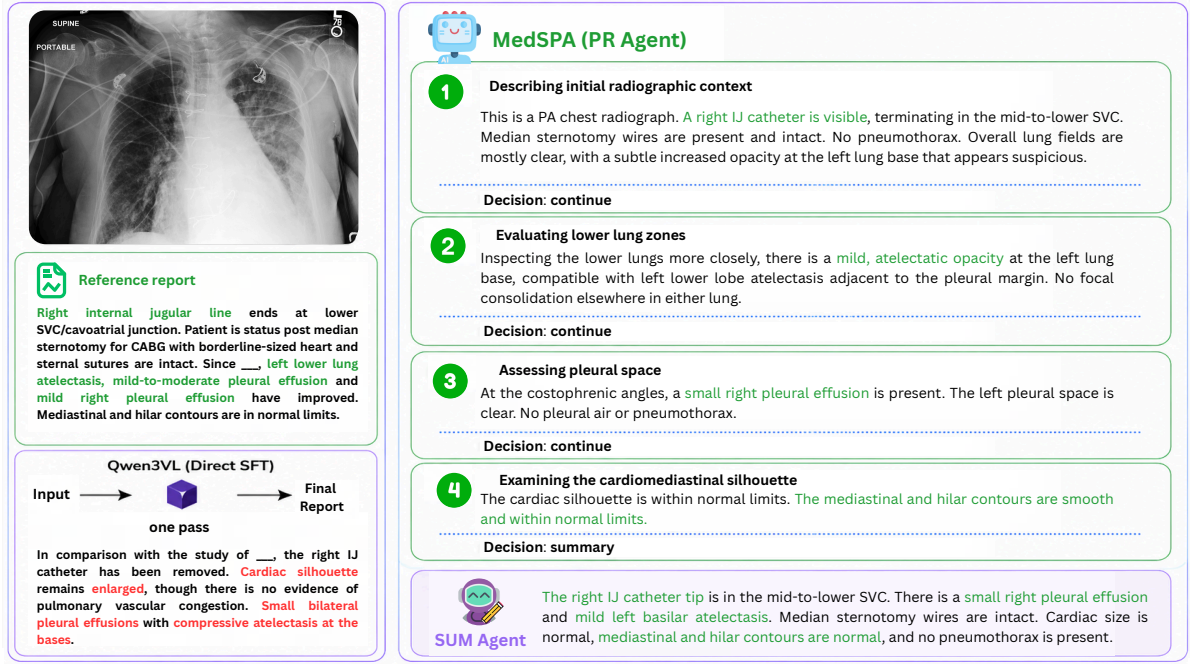


Figure 3: Qualitative analysis of MedSPA on MIMIC-CXR report generation.

Table 2: Comparison of base MLLMs and MedSPA variants on the MedVQA task. O denotes the F1 score for open-ended questions, and C denotes the accuracy for closed-ended questions. Higher values indicate better performance, and the best score in each column is shown in bold.

Method	SLAKE								VQA-RAD							
	Chest X-ray		MRI Brain		CT Abdomen		Average		Chest X-ray		MRI Brain		CT Abdomen		Average	
	O	C	O	C	O	C	O-avg	C-avg	O	C	O	C	O	C	O-avg	C-avg
Gemini 2.5 Pro (Gemini Team, 2025)	52.3	85.7	44.5	71.7	54.9	76.4	50.6	77.9	35.2	71.6	50.3	80.0	35.3	70.7	40.3	74.1
LLaVA (Liu et al., 2023)	35.4	78.6	30.2	83.2	41.1	73.9	35.6	78.6	26.8	59.6	45.7	65.4	24.9	69.1	32.5	64.7
MedSPA-LLaVA	60.9	79.6	54.8	84.2	62.3	75.8	59.3	79.9	30.9	67.8	50.6	72.9	25.4	74.5	35.6	71.7
Qwen3VL (Bai et al., 2025)	38.7	80.7	32.0	78.9	44.1	75.2	38.3	78.3	28.7	60.6	48.7	66.7	21.2	70.8	32.9	66.0
MedSPA-Qwen3VL	81.8	85.1	70.2	80.6	70.5	77.4	74.2	81.0	32.6	75.2	59.7	83.3	26.3	71.9	39.5	76.8

backbone (Qwen3-VL-8B-Instruct) across SLAKE and VQA-RAD. The improvement is especially pronounced for open-ended questions, where the model must generate free-form answers from visual evidence rather than select from options. On SLAKE, MedSPA-Qwen3VL improves the open-ended average of Qwen3VL by 35.9 points, while MedSPA-LLaVA improves LLaVA by 23.7 points. On VQA-RAD, MedSPA also improves the closed-ended average for both backbones and achieves the best overall closed-ended accuracy with Qwen3VL. These results suggest that self-planned evidence acquisition is particularly useful when the model must identify visual evidence before producing an answer.

Overall, the results demonstrate that MedSPA improves both long-form report generation and question-specific medical reasoning. The consistent gains over Direct SFT and other baselines sup-

port our central claim that planning diagnostic evidence before prediction leads to more clinically grounded medical outputs.

4.3 Ablation Studies

We analyze the contribution of MedSPA components and training stages in Table 3. PR and SUM in the table mean whether the PR and SUM agents after supervised fine-tuning are used or not. Since the PR agent does not generate a final report, we evaluate *PR only* using its accumulated reasoning memory, which contains predicted labels of findings and is compared with ground-truth findings extracted from the reference report. The PR agent clearly improves the baseline (Direct SFT), especially increasing recall by 0.084, indicating that the PR agent expands evidence coverage by encouraging the model to explore a broader set of clinically relevant findings before generating the final output.

Table 3: Ablation study of MedSPA components on the MIMIC-CXR test set. PR-only variants are evaluated using reasoning memory labels, while variants with SUM are evaluated using the final generated reports.

Model	PR	SUM	R-Align	Prec.	Rec.	F1
Direct SFT	–	–	–	0.483	0.466	0.474
PR only	✓	–	–	0.540	0.550	0.545
SFT MedSPA	✓	✓	–	0.549	0.543	0.546
R-Aligned PR	✓	–	✓	0.549	0.563	0.556
R-Aligned MedSPA	✓	✓	✓	0.569	0.552	0.560

Incorporating the SUM agent improves precision by 0.009, potentially suggesting that irrelevant evidence collected by the PR agent is filtered by the SUM agent.

In MedSPA, broader evidence exploration introduces a precision-recall trade-off: the PR agent increases the chance of discovering useful findings and improves recall, thereby providing richer evidence for the SUM agent to improve final prediction precision, but it may also introduce uncertain or weakly supported evidence. Therefore, the goal of MedSPA is not to simply explore more evidence, but to guide the model toward the right diagnostic directions and stop before unnecessary exploration degrades prediction precision. Reinforcement alignment further improves this balance, with R-Aligned MedSPA improving F1 by 0.086 over the baseline.

4.4 Reasoning Strategy Analysis

To investigate the impact of reasoning strategy on MRG, we conduct controlled zero-shot evaluations on MIMIC-CXR using three representative vision-language models: DeepSeek-VL2 (Wu et al., 2024), Llama-4 Scout (Meta AI, 2025), and Gemini 2.5 Pro (Gemini Team, 2025). For each model, we compare three reasoning settings: chain-of-thought (CoT), fixed template, and MedSPA planning (PR plan). The fixed template uses a predefined 10-step medical checklist,⁷ while MedSPA planning uses diagnostic planning actions generated by the PR agent to provide case-specific guidance for subsequent reasoning.

Figure 4 shows that the reasoning strategy itself has a clear impact on clinical performance. Compared with CoT, the fixed template generally improves recall by enforcing more complete diagnostic coverage, but it requires the full 10-step traversal regardless of case complexity. In contrast,

⁷Details of the fixed medical templates are provided in the Appendix.

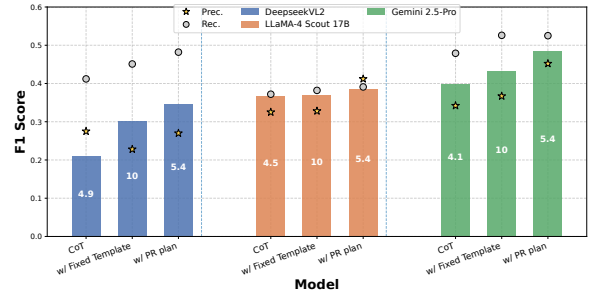


Figure 4: Impact of different reasoning strategies on MIMIC-CXR report generation performance across multiple VLLMs. Numbers inside the bars indicate the average number of reasoning steps.

MedSPA planning achieves higher F1 scores across different VLLMs while using fewer reasoning steps than the fixed template. CoT typically produces around 4–5 reasoning steps, the fixed template follows 10 predefined steps, and MedSPA planning uses approximately five steps on average. This suggests that the benefit of MedSPA does not come from simply increasing reasoning length, but from guiding the model toward case-aware diagnostic evidence. These results support our design choice of leveraging clinical templates as soft planning priors and converting them into adaptive, MedSPA-guided evidence acquisition rather than rigid template traversal.

4.5 Further Analysis

Additional Clinical Evaluation. The current R-Align reward design relies on CheXpert-derived clinical labels, while the primary CE-F1 evaluation is based on the same predefined label space. This raises concerns about potential evaluation bias toward improvements within the same predefined label space. We therefore additionally evaluate SFT MedSPA and R-Aligned MedSPA using GREEN (Ostmeier et al., 2024) and RadCliQ (Yu et al., 2023). GREEN evaluates clinically significant report errors beyond the predefined CheXpert label space, with higher scores indicating better clinical agreement, while RadCliQ provides a broader report-level estimate of clinical error, where lower scores indicate fewer errors. Table 4 shows that MedSPA achieves the best performance among the compared methods on both RadCliQ and GREEN. Importantly, R-Align further improves the SFT model on both metrics, increasing GREEN by 0.005 and reducing RadCliQ by 0.021. These results suggest that R-Align’s benefits extend beyond the CheXpert-based CE-F1 criterion.

Model	GREEN $\uparrow$	RadCliQ $\downarrow$
<i>General-purpose MLLMs</i>		
GPT-4V	0.160	1.786
GPT-4o	–	1.493
GPT-5	–	1.592
Gemini 2.5 Pro	–	1.543
<i>CXR-specific models</i>		
R2Gen	0.276	1.228
PromptMRG	0.287	1.169
<i>Our models</i>		
SFT MedSPA	0.290	0.682
R-Aligned MedSPA	0.295	0.661

Table 4: Additional clinical evaluation on the MIMIC-CXR test set using GREEN and RadCliQ. Higher GREEN and lower RadCliQ indicate better clinical report quality.

	Gemini 2.5 Pro	GPT-4o
Reasoning Chains	4.4 / 5	4.2 / 5
Generated Reports	4.5 / 5	4.1 / 5

Table 5: LLM-as-a-judge evaluation of MedSPA on 100 randomly sampled MIMIC-CXR test cases.

LLM-as-a-Judge Evaluation. Beyond these clinical metrics, we further evaluate MedSPA’s reasoning chains and generated reports on 100 randomly sampled MIMIC-CXR test cases using two independent judges, Gemini 2.5 Pro (Gemini Team, 2025) and GPT-4o (OpenAI, 2024). For each case, the judges separately assess the clinical correctness and completeness of the reasoning chain generated by the PR Agent and the final report generated by the SUM Agent with respect to the corresponding reference report. Each output is scored on a 1–5 scale.⁸ As shown in Table 5, both judges assign average scores above 4.0 to MedSPA’s reasoning chains and generated reports, indicating that both outputs generally preserve clinically relevant information and remain consistent with the references.

Inference Efficiency. We also analyze the inference overhead introduced by the iterative PR–SUM pipeline, highlighting the trade-off between improved clinical performance and increased inference cost. As shown in Table 6, MedSPA performs 5.4 PR steps followed by one SUM generation, incurring an additional 2.04 seconds of latency over Direct SFT while improving CE-F1 by 0.086. This additional cost mainly arises from the explicit evidence-acquisition process, in which the

Model	Calls	Latency $\downarrow$	CE-F1 $\uparrow$
Direct SFT	1.0	0.41 s	0.474
MedSPA	6.4	2.45 s	0.560

Table 6: Inference efficiency on the MIMIC-CXR test set. Latency denotes the average end-to-end inference time per sample.

PR Agent sequentially determines what diagnostic aspect to inspect, updates its reasoning memory, and decides whether further evidence is still needed before invoking the SUM Agent. Overall, these results show that MedSPA trades additional inference time for stronger clinical performance and more controllable diagnostic reasoning.

5 Conclusion

This study investigates how medical vision-language models can perform reliable diagnostic reasoning by planning evidence acquisition before prediction with our two-agent framework, MedSPA. Experiments on medical report generation and medical visual question answering show that MedSPA improves clinical faithfulness, controllability, and diagnostic reliability. These results support our central view that reliable medical reasoning requires models to acquire clinically relevant evidence in a planned, case-adaptive manner and to stop before unnecessary exploration weakens prediction precision. Future work can extend self-planned reasoning to broader clinical scenarios, incorporate richer expert feedback, improve curated reasoning-chain quality, and develop more fine-grained evaluations for reasoning faithfulness.

Limitations

A limitation of our work is that MedSPA lags behind some previous report-generation methods on standard NLG metrics. For example, DART (Park et al., 2025) obtains higher BLEU, ROUGE-L, and METEOR scores on MIMIC-CXR. This gap may partly arise from differences in training objectives: DART is optimized through direct report-level supervision, which more closely aligns with lexical-overlap metrics, whereas MedSPA prioritizes the acquisition and preservation of clinically relevant evidence before report generation. Consequently, clinically equivalent findings may be expressed using wording that differs from the reference report, resulting in lower lexical-overlap scores. We provide a qualitative example of this discrepancy in

⁸Details of the evaluation protocol are provided in Appendix F.2.

Appendix H. Moreover, our current evaluation does not directly assess report fluency and coherence through human evaluation. In future work, we will investigate language-aware reward functions and expert-based evaluation to better balance clinical fidelity with linguistic quality.

A second issue concerns the fact that the current R-Align reward design operates within the 14-label CheXpert space. Consequently, clinically relevant findings outside this predefined label set, such as device positioning, skeletal abnormalities, and subdiaphragmatic free air, are not directly reinforced during R-Align. However, these findings remain part of the supervision during SFT through the complete reasoning chains and reference reports. We further examine this issue using GREEN and RadCliQ in Table 4, where MedSPA maintains strong report-level clinical performance beyond the CheXpert-based CE-F1 evaluation. Nevertheless, explicitly optimizing findings outside the predefined label space requires further investigation.

Another consideration is that the current PR Agent does not perform explicit ROI-guided visual planning, visual grounding, or attention control, raising questions about whether the model attends to the image regions relevant to each reasoning step. In this work, we focus on the planning capability of an MLLM over the entire medical image, where each planning action, such as “Inspect the lungs” or “Assess the pleural space” serves as a language instruction that conditions the model’s subsequent visual reasoning over the same chest X-ray. Prior studies have shown that modern vision-language models can exhibit implicit visual grounding, in which language instructions guide attention toward semantically relevant image regions without requiring explicit ROI supervision or additional grounding modules (Kang et al., 2025; Wang et al., 2025). Thus, although MedSPA does not explicitly localize a region before each reasoning step, its planning actions may still guide the model toward the corresponding clinical evidence. Nevertheless, incorporating ROI-aware planning represents a promising direction for improving evidence localization and interpretability.

Another challenge lies in verifying the clinical correctness of every intermediate reasoning step in the training data used for MedSPA. The current curation pipeline cannot fully establish the correctness of each generated step, while exhaustive expert review and correction are difficult to scale because each training sample typically con-

tains approximately five to six reasoning steps. As a practical mitigation, we employ the quality-control pipeline described in Section 3.2, which combines clinical label-consistency checking with LLM-based assessment of faithfulness and completeness to reduce unsupported or incomplete reasoning before SFT. We further present a representative failure case in Appendix G.

Regarding teacher selection, Qwen3VL-32B is not the largest or most capable vision-language model currently available, and stronger teacher models may produce higher-quality reasoning chains and further improve downstream student performance. Our teacher sensitivity analysis in Appendix B.2 supports this possibility, showing that student performance improves across the teacher models considered in our study. Although Qwen3VL-32B achieves reasoning-chain quality comparable to Gemini 2.5 Pro and offers a practical trade-off between reasoning quality and large-scale curation cost, we do not exhaustively evaluate larger or more capable teacher models. Therefore, the extent to which stronger teachers could further improve MedSPA remains to be explored.

Ethical Considerations

We conduct all experiments using existing de-identified medical imaging datasets accessed under their respective data-use requirements and do not collect new patient data. The proposed model is intended for research purposes only. Despite its improvements in clinical faithfulness, it may still generate incorrect or incomplete medical information and should not be used for clinical diagnosis or decision-making without appropriate validation and human oversight.

Acknowledgments

This work was supported by JST ASPIRE Grant No. JPMJAP2502, JST FOREST Grant No. JPMJFR2160, JSPS KAKENHI Grant No. JP25K24609 and JP23H00497, JST ACT-X Grant No. JPMJAX24C8.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhi-fang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng

- Li, and 45 others. 2025. Qwen3-VL technical report. *Preprint*, arXiv:2511.21631.
- Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.
- Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, Mercy Ranjit, Shaurya Srivastav, Julia Gong, Noel C. F. Codella, Fabian Falck, Ozan Oktay, Matthew P. Lungren, Maria Teodora Wetscherek, and 2 others. 2024. MAIRA-2: Grounded radiology report generation. *Preprint*, arXiv:2406.04449.
- Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2021. Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 5904–5914.
- Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 1439–1449.
- Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.
- Sedigheh Eslami, Christoph Meinel, and Gerard de Melo. 2023. PubMedCLIP: How much does CLIP benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1181–1193.
- Arsene Fansi Tchango, Rishab Goel, Julien Martel, Zhi Wen, Gaetan Marceau Caron, and Joumana Ghosn. 2022. Towards trustworthy automatic diagnosis systems by emulating doctors’ reasoning with deep reinforcement learning. *Advances in Neural Information Processing Systems*, 35:24502–24515.
- Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next-generation agentic capabilities. *Preprint*, arXiv:2507.06261.
- William Herring. 2019. *Learning Radiology: Recognizing the Basics*, 4th edition. Elsevier.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.
- Marc Sebastian Huppertz, Robert Siepmann, David Topp, Omid Nikoubashman, Can Yüksel, Christiane Katharina Kuhl, Daniel Truhn, and Sven Nebelung. 2025. Revolution or risk?—assessing the potential and challenges of GPT-4V in radiologic image interpretation. *European Radiology*, 35(3):1111–1121.
- Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597.
- 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. 2021. RadGraph: Extracting clinical entities and relations from radiology reports. *Preprint*, arXiv:2106.14463.
- Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. 2024. PromptMRG: Diagnosis-driven prompts for medical report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2607–2615.
- Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. 2019. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *Preprint*, arXiv:1901.07042.
- Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. 2025. Your large vision-language model only needs a few attention heads for visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9339–9350.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35:22199–22213.
- Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, Yuheng Li, Konstantinos Psounis, and Xiaofeng Yang. 2026. Med-R1: Reinforcement learning for generalizable medical reasoning in vision-language models. *IEEE Transactions on Medical Imaging*, 45(6):2727–2737.
- Jason J. Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5:180251.

- Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.
- Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.
- Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *Proceedings of the IEEE International Symposium on Biomedical Imaging*, pages 1650–1654.
- Chang Liu, Yuanhe Tian, Weidong Chen, Yan Song, and Yongdong Zhang. 2024a. Bootstrapping large language models for radiology report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18635–18643.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916.
- Jiaxiang Liu, Yuan Wang, Jiawei Du, Joey Tianyi Zhou, and Zuozhu Liu. 2024b. MedCoT: Medical chain of thought via hierarchical expert. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17371–17389.
- Rui Liu, Mingjie Li, Shen Zhao, Ling Chen, Xiaojun Chang, and Lina Yao. 2024c. In-context learning for zero-shot medical report generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8721–8730.
- Meta AI. 2025. The Llama 4 herd: The beginning of a new era of natively multimodal ai innovation.
- Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fiedland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharmashan Kumar, Daan Wierstra, Shane Legg, and Demis Hassabis. 2015. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.
- Johannes Moll, Markus Graf, Tristan Lemke, Nicolas Lenhart, Daniel Truhn, Jean-Benoit Delbrouck, Jiazheng Pan, Daniel Rueckert, Lisa C. Adams, and Keno K. Bresssem. 2026. Evaluating reasoning faithfulness in medical vision-language models using multimodal perturbations. In *Proceedings of the Fifth Machine Learning for Health Symposium*, volume 297 of *Proceedings of Machine Learning Research*, pages 424–448. PMLR.
- OpenAI. 2023. GPT-4V(ision) system card.
- OpenAI. 2024. GPT-4o system card. *Preprint*, arXiv:2410.21276.
- 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. 2024. GREEN: Generative radiology report evaluation and error notation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 374–390.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.
- Sang-Jun Park, Keun-Soo Heo, Dong-Hee Shin, Young-Han Son, Ji-Hye Oh, and Tae-Eui Kam. 2025. 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*, pages 15580–15589.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741.
- Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, Juanma Zambano Chaves, Szu-Yeu Hu, Mike Schaekermann, Aishwarya Kamath, Yong Cheng, David G. T. Barrett, Cathy Cheung, Basil Mustafa, Anil Palepu, and 48 others. 2024. Capabilities of Gemini models in medicine. *Preprint*, arXiv:2404.18416.
- Hongyu Shen, Mingtao Pei, Juncai Liu, and Zhaoxing Tian. 2024. Automatic radiology reports generation via memory alignment network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4776–4783.
- Harjit Singh, Janet A. Neutze, and Jonathan R. Enterline, editors. 2015. *Radiology Fundamentals: Introduction to Imaging & Technology*, 5th edition. Springer.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguerre y Arcas, Dale Webster, and 11 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.
- Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. 2023. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7433–7442.

- Bowen Wang, Jiuyang Chang, Yiming Qian, Guoxin Chen, Junhao Chen, Zhouqiang Jiang, Jiahao Zhang, Yuta Nakashima, and Hajime Nagahara. 2024. Direct: Diagnostic reasoning for clinical notes via large language models. *Advances in neural information processing systems*, 37:74999–75011.
- Shansong Wang, Mingzhe Hu, Qiang Li, Mojtaba Safari, Richard L. J. Qiu, and Xiaofeng Yang. 2026. Capabilities of GPT-5 on multimodal medical reasoning. In *Medical Imaging 2026: Imaging Informatics*, volume 13930, pages 77–86.
- Shijie Wang, Dahun Kim, Ali Taalimi, Chen Sun, and Weicheng Kuo. 2025. Learning visual grounding from generative vision and language model. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8057–8067.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023a. Self-consistency improves Chain-of-Thought reasoning in language models. In *International Conference on Learning Representations*.
- Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2023b. METransformer: Radiology report generation by transformer with multiple learnable expert tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11558–11567.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.
- Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, and 8 others. 2024. DeepSeek-VL2: Mixture-of-experts vision-language models for advanced multimodal understanding. *Preprint*, arXiv:2412.10302.
- Peng Xia, Ze Chen, Juanxi Tian, Yangrui Gong, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, Wenhao Zheng, Zhaoyang Wang, Xiao Wang, Xuchao Zhang, Chetan Bansal, Marc Niethammer, Junzhou Huang, Hongtu Zhu, Yun Li, and 5 others. 2024. CARES: A comprehensive benchmark of trustworthiness in medical vision-language models. *Advances in Neural Information Processing Systems*, 37:140334–140365.
- Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. 2023. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9):100802.
- Jaehoon Yun, Jiwoong Sohn, Jungwoo Park, Hyunjae Kim, Xiangru Tang, Daniel Shao, Yonghoe Koo, Minhyeok Ko, Qingyu Chen, Mark Gerstein, Michael Moor, and Jaewoo Kang. 2025. Med-PRM: Medical reasoning models with stepwise, guideline-verified process rewards. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16565–16582.
- Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. 2023. HuatuoGPT: Towards taming language model to be a doctor. In *Findings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10859–10885.
- Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D. Davison, Hui Ren, Jing Huang, Chen Chen, Yuyin Zhou, Sunyang Fu, Wei Liu, Tianming Liu, Xiang Li, Yong Chen, Lifang He, and 4 others. 2024. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, 30(11):3129–3141.
- Kangyu Zhu, Peng Xia, Yun Li, Hongtu Zhu, Sheng Wang, and Huaxiu Yao. 2024. MMedPO: Aligning medical vision-language models with clinical-aware multimodal preference optimization. *Preprint*, arXiv:2412.06141.
- Zhihong Zhu, Yunyan Zhang, Xianwei Zhuang, Fan Zhang, Zhongwei Wan, Yuyan Chen, Qingqing Long, Yefeng Zheng, and Xian Wu. 2025. Can we trust AI doctors? a survey of medical hallucination in large language and large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6748–6769. Association for Computational Linguistics.

Appendix

A Details of Inference

A.1 PR Agent Inference Prompt

At inference time, the PR Agent iteratively generates one reasoning step at a time.

Note: Text in brackets is shown only for readability and is not included in the actual prompt.

PR Agent Inference Prompt

```
<image>
You are a highly experienced radiologist.
[If there is no previous reasoning step]
There are no previous reasoning steps yet.
Generate only the first reasoning step based on the image.
[Otherwise]
Here are the previous reasoning steps:

Step 1: {previous reasoning step 1}
Step 2: {previous reasoning step 2}
...
Step t-1: {previous reasoning step t-1}

Generate only the next reasoning step based on the image
and the previous steps above.
```

A.2 SUM Agent Inference Prompt

After the PR Agent terminates the evidence-acquisition loop, the SUM Agent generates the final radiology report or answer.

SUM Agent Inference Prompt

```
<image>
You are a highly experienced radiologist.
Here are the clinical reasoning steps for this image:

Step 1: {reasoning step 1}
Step 2: {reasoning step 2}
...
Step T: {reasoning step T}

[For MRG]
Based on the provided image and the above reasoning
steps, write a clear, clinically appropriate radiology report.
[For MedVQA]
Based on the provided image and the above reasoning
steps, answer the following clinical question based on
the image. Be concise and clinically accurate. Question:
{question}
```

B Reasoning Chain Curation Prompt

We construct the reasoning-chain curation prompt using modular instruction components.

1. General Role and Goal

You are a highly experienced radiologist that generates a clear, step-wise interpretation of a medical image. Your goal is to produce a concise, clinically grounded reasoning chain that:

- is consistent with the underlying radiographic truth,
- may internally use the provided findings and report

only as hidden guidance,

- remains strictly image-grounded and medically consistent in its wording.

2. Soft Medical Reasoning Template

This soft medical template describes how to analyze a medical image step by step.

1. First step: global context

- The first step must describe the initial radiographic context.
- The first step should mention the image type and projection, if identifiable.
- It should also mention the presence or absence of obvious support devices, lines, or tubes if visible.
- Do not jump directly into a single abnormality; this step sets the scene.

2. Then explore major regions in a natural, radiologist-like order

- You do not have to follow the textual order of the trusted findings, as long as you remain consistent.
- When examining a region, first describe the overall status and important negative findings, then describe positive or uncertain abnormalities.
- Normal areas are described first, while abnormalities are described later.

3. Step structure and focus

- Each step should focus on one coherent idea, anatomical region, or diagnostic aspect.
- Avoid mechanically scanning every possible region if it is not relevant to the trusted findings.
- Later steps must not contradict earlier steps.
- Every step title should be a short V-ing phrase summarizing the purpose of the step.

4. Final step

- When the collected evidence is sufficient to support the main radiographic findings, conclude the reasoning chain by setting the final decision to "summary".
- It must remain image-grounded and consistent with the trusted findings.

We use RadGraph (Jain et al., 2021) to extract key medical entities and their relations from the original radiology report. These extracted entities and relations are converted into a concise list of radiographic findings, which is used as trusted semantic guidance during chain generation in component 3 below.

3. Trusted Radiographic Findings

You will receive a list of trusted findings derived from an expert radiology report.

- These findings represent the most reliable judgment about what is present, absent, or uncertain.
- Treat the trusted findings as the primary source of truth about radiographic content.
- Do not invent new radiographic abnormalities that are not in the trusted findings or not clearly visible in the image.

Trusted Findings: {Trusted Findings}

4. Report Guidance

You will also receive the original radiology report as a knowledge source. Treat the report solely as internal clinical context and hypothesis cues that assist image interpretation. Do not introduce new radiographic abnormalities that are neither in the trusted findings nor clearly visible in the image. Report: {Report}

5. Output Format

Return JSON object with exactly this structure:

```
[
  {
    "action": "short V-ing phrase summarizing the purpose of the step",
    "reason": "objective image-grounded reasoning",
    "decision": "continue" or "summarize"
  },
  ...
]
```

B.1 Quality Control

We use an LLM evaluator to score whether each generated reasoning chain is faithful to the reference caption and sufficiently covers the clinically salient findings. The prompt used for quality scoring is shown below.

LLM Reasoning Chain Quality Scorer

You are a strict radiology quality-assurance scorer with clinical expertise. Grade the entire reasoning chain against the reference report/answer.

Reference report: {report}

Reasoning chain: {Reasoning chain}

Scoring: Assign an integer score from 0 to 100. Use equal importance for the following criteria:

1. **Faithfulness:** factual alignment with the reference caption. Penalize hallucinations, contradictions, wrong laterality or location, and unsupported claims.
2. **Completeness:** coverage of the salient elements that the chain should address for this case, including key findings, location or laterality, extent or severity, and relevant devices or context. Penalize missing critical details or vague coverage. Do NOT reward

unsupported content.

Output rules:

- Output exactly one integer between 0 and 100 on a single line.
- Do not output words, units, JSON, punctuation, or explanations.

B.2 Teacher Model Analysis

Teacher Model Selection To select a teacher model for reasoning chain curation, we evaluate several candidate vision-language models on 5,000 randomly sampled cases from the MIMIC-CXR training split. For each model, we generate reasoning chains using the same curation prompt and score the generated chains with the LLM reasoning-chain quality scorer described in Section B.1. The score measures whether the generated chain is faithful to the reference caption and sufficiently covers clinically salient findings.

As shown in Table 10, Gemini 2.5 Pro and Qwen3VL-32B achieve the highest reasoning-chain quality score. Given the comparable quality between Gemini 2.5 Pro and Qwen3VL-32B, and considering the large scale of our curation pipeline, we select Qwen3VL-32B as the teacher model for the final dataset construction because it provides a better trade-off between reasoning quality and curation cost.

Tables 8 and 9 summarize the statistics of the curated reasoning-chain datasets for medical report generation and medical visual question answering. Representative samples from the curation data pipeline are shown in Figure 5.

Teacher Sensitivity. We further examine how the capability of the teacher used for reasoning-chain generation affects downstream student performance. We use Qwen2-VL-7B-Instruct, Qwen3-VL-8B-Instruct, and Qwen3-VL-32B-Instruct to generate approximately 120K reasoning chains from the MIMIC-CXR training set. To isolate the effect of teacher choice, we keep the student backbone (Qwen2.5-VL) and training setting fixed across all experiments and train each student for three epochs.

As shown in Table 7, student performance improves across the teacher models considered in this study, with CE-F1 increasing from 0.432 to 0.450 and 0.468. These results indicate that teacher capability influences the quality of reasoning supervision and downstream student performance, sug-

Table 7: Effect of teacher capability on MedSPA performance on MIMIC-CXR.

Teacher Model	Student	Precision $\uparrow$	Recall $\uparrow$	CE-F1 $\uparrow$
Qwen2-VL-7B	Qwen2.5-VL	0.441	0.426	0.432
Qwen3-VL-8B	Qwen2.5-VL	0.458	0.445	0.450
Qwen3-VL-32B	Qwen2.5-VL	0.476	0.462	0.468

Table 8: Statistics of the curated dataset for medical report generation distillation.

Statistic	Value
Number of distilled samples	268821
Average reasoning steps	5.4
Minimum reasoning steps	3
Maximum reasoning steps	8

gesting that stronger teachers may further improve MedSPA.

C Additional Details of R-Align

C.1 SUM Label-Level Reward

For each clinical label $l \in \mathcal{L}$, let $f_l \in \{0, 1\}$ denote whether label l appears in the reference annotation, $v_l^R \in \{0, 1\}$ denote whether it is extracted from the PR Agent’s reasoning trajectory, and $v_l^S \in \{0, 1\}$ denote whether it is extracted from the SUM Agent’s final output. The label-level reward $r(v_l^R, v_l^S, f_l)$ used in Eq. (5) is defined in Table 11.

C.2 Optimization Details of Role-Specific Reinforcement Alignment

This section provides the full optimization objectives used in R-Align.

C.2.1 PR Agent Objective

For each input, the PR Agent samples N reasoning trajectories $\{Z_i\}_{i=1}^N$ and computes the role-specific reward $R^R(Z_i)$ as defined in Section 3.4. The group-relative advantage for the PR Agent is:

$$A_i^R = R^R(Z_i) - \frac{1}{N} \sum_{j=1}^N R^R(Z_j). \quad (6)$$

Let π_θ^R denote the current PR policy and π_{ref}^R denote the SFT reference policy. The PR Agent is optimized with a KL-regularized group-relative

Table 9: Statistics of the curated dataset for medical visual question answering distillation.

Dataset	Subset	# Open	# Closed	# Total	Avg. Steps
SLAKE	Chest X-ray	969	454	1,423	5.2
	CT Brain	196	99	295	5.6
	CT Abdomen	716	599	1,315	5.3
VQA-RAD	Chest X-ray	241	355	596	5.1
	Brain	44	41	85	5.8
	Abdomen	264	310	574	5.5

Table 10: Teacher model comparison for reasoning chain distillation on 5,000 MIMIC-CXR training samples. Scores are computed by the LLM reasoning-chain quality scorer.

Model	Quality Score	Avg. Steps
DeepSeek-VL2	70	4.9
Llama4-Scout-17B	75	4.5
Gemini 2.5 Pro	85	4.1
Qwen3VL-32B	85	5.4

objective:

$$\mathcal{L}_R = -\mathbb{E}_{Z_i \sim \pi_\theta^R} \left[A_i^R \sum_{t=1}^{T_i} \log \pi_\theta^R(z_{i,t} \mid I, q, z_{i,<t}) \right] + \lambda_{\text{KL}} \text{KL}(\pi_\theta^R \parallel \pi_{\text{ref}}^R), \quad (7)$$

where λ_{KL} controls the strength of Kullback-Leibler regularization.

C.2.2 SUM Agent Objective

After freezing the PR Agent, the SUM Agent samples M candidate outputs $\{Y_i\}_{i=1}^M$ and computes the summary reward $R^S(Y_i)$ as defined in Section 3.4. The group-relative advantage for the SUM Agent is:

$$A_i^S = R^S(Y_i) - \frac{1}{M} \sum_{j=1}^M R^S(Y_j). \quad (8)$$

Let π_θ^S denote the current SUM policy and π_{ref}^S denote the SFT reference policy. The SUM Agent is optimized with:

$$\mathcal{L}_S = -\mathbb{E}_{Y_i \sim \pi_\theta^S} [A_i^S \log \pi_\theta^S(Y_i \mid I, q, Z)] + \lambda_{\text{KL}} \text{KL}(\pi_\theta^S \parallel \pi_{\text{ref}}^S). \quad (9)$$

Both objectives use the SFT model as the reference policy to prevent the aligned agents from drifting too far from the curated reasoning behavior learned during supervised fine-tuning.



GT Report: There moderate pulmonary edema with perihilar opacities as well as vascular indistinctness. There are likely small pleural effusions bilaterally. The heart size is enlarged. Calcifications of aorta are noted. Patchy opacities in the lung bases likely reflect atelectasis though infection cannot be excluded in the correct clinical setting. No pneumothorax. A chronic appearing deformity of the right clavicle involving resorption/resection of the distal aspect is noted along with severe degenerative changes of the right glenohumeral joint.

Reasoning Step

1. Action: Describing initial radiographic context

Reason: This is an upright chest radiograph. No endotracheal tube, chest tube, or other acute foreign body is visible. Overall image quality is adequate for assessment of the lungs, cardiomeastinal silhouette, pleural spaces, and osseous structures.

Decision: continue

2. Title: Assessing cardiopulmonary findings

Reason: The cardiac silhouette is enlarged, consistent with cardiomegaly. There is moderate pulmonary edema with bilateral perihilar opacities and vascular indistinctness. Patchy opacities are present at the lung bases, likely reflecting atelectasis, although infection cannot be excluded in the appropriate clinical setting.

Decision: continue

3. Title: Evaluating pleural spaces and pneumothorax

Reason: Mild bilateral costophrenic angle blunting suggests small bilateral pleural effusions. No pneumothorax is identified, as there is no visible pleural line or peripheral lucency, and lung markings extend to the chest wall.

Decision: continue

4. Title: Reviewing mediastinal and aortic structures

Reason: The mediastinum is not widened or shifted. Atherosclerotic calcification is visible along the thoracic aorta. No obvious acute mediastinal abnormality is seen on this radiograph.

Decision: continue

5. Title: Examining osseous structures and chronic joint changes

Reason: There is a chronic-appearing deformity of the right clavicle involving resorption/resection of the distal aspect. Severe degenerative changes of the right glenohumeral joint are present.

Decision: summary

Figure 5: Sample of curated plan-reasoning data for MIMIC-CXR report generation

Table 11: Binary cases for the SUM label-level reward.

f_l	v_l^R	v_l^S	$r(v_l^R, v_l^S, f_l)$
1	1	1	+1
1	1	0	-1
1	0	1	0
1	0	0	0
0	1	1	-1
0	1	0	0
0	0	1	-1
0	0	0	0

D Additional Training Details

D.1 Supervised Fine-Tuning

We first train the PR and SUM agents using supervised fine-tuning on the curated reasoning-chain dataset. The main hyperparameters for supervised fine-tuning are:

- Backbone model: Qwen3-VL-8B-Instruct.
- Fine-tuning method: LoRA with rank $r = 64$, $\alpha = 128$, and dropout 0.05.
- Optimizer: AdamW.
- Training epochs: 10 for MRG and 3 for Med-VQA

- Learning rate: 2×10^{-5} .
- Weight decay: 0.05.
- Maximum sequence length: 4096 tokens.
- Image size: 1024×1024 .

D.2 Role-Specific Reinforcement Alignment

After supervised fine-tuning, we apply Role-specific reinforcement alignment to further optimize the PR and SUM agents. The main hyperparameters for R-Align are:

- Initialization: SFT PR Agent, SFT SUM Agent.
- Trainable adapter: separate GRPO LoRA adapter.
- R-Align training samples: 50,000.
- R-Aligned PR Agent training epochs: 2.
- R-Aligned SUM Agent training epochs: 2.
- Maximum prompt length: 512 tokens.
- Maximum sequence length: 4096 tokens.
- Learning rate: 2×10^{-5} .
- Coverage penalty coefficient β : 0.5.

E Fixed Medical Template for Reasoning Strategy Analysis

This section describes the fixed medical template used in the reasoning strategy analysis. The model must complete all steps in the given order before generating the final output, regardless of case complexity or whether sufficient evidence has already been collected.

Fixed 10-Step Medical Template

You must analyze the chest X-ray by strictly following the 10 diagnostic steps below. Do not skip, reorder, merge, or stop before completing all steps. After finishing Step 10, generate the final report or answer based on the observations from all steps.

- Initial radiographic context.** Identify the image type, projection/view, image quality, and overall thoracic appearance.
- Support devices and postsurgical changes.** Check for tubes, lines, catheters, pacemakers, sternotomy wires, surgical clips, or other devices.
- Airway and mediastinal alignment.** Examine the trachea, mediastinal position, and any shift or widening.
- Cardiomediastinal silhouette.** Assess heart size, cardiac contour, mediastinal contour, and hilar structures.
- Right lung field.** Inspect the right upper, middle, and lower lung zones for opacity, atelectasis, consolidation, edema, or nodules.
- Left lung field.** Inspect the left upper, middle, and lower lung zones for opacity, atelectasis, consolidation, edema, or nodules.
- Pleural spaces.** Assess both pleural spaces and costophrenic angles for pleural effusion, pleural thickening, or pneumothorax.
- Pulmonary vascularity.** Evaluate pulmonary vascular congestion, interstitial edema, and signs of volume overload.
- Bones and soft tissues.** Examine visible osseous structures and chest wall soft tissues for acute abnormality.
- Diaphragm and upper abdomen.** Inspect the hemidiaphragms, subdiaphragmatic regions, and visible upper abdomen for free air, abnormal elevation, or other acute findings.

F Additional Evaluation Results

F.1 Comparison with Strong General-purpose MLLMs

We additionally evaluate several strong general-purpose MLLMs on the MIMIC-CXR test set, including DeepSeek-VL2, Llama-4-Scout-17B-16E,

Model	Prec. ↑	Rec. ↑	F1 ↑	B-1 ↑	B-4 ↑	R-L ↑	M ↑
DeepSeek-VL2	0.275	0.171	0.211	0.231	0.015	0.166	0.087
Llama-4-Scout-17B-16E	0.325	0.372	0.367	0.154	0.009	0.143	0.124
GPT-4V	0.353	0.365	0.330	0.338	0.061	0.240	0.125
Gemini 2.5 Pro	0.342	0.479	0.399	0.232	0.028	0.182	0.137
SFT MedSPA	0.549	0.543	0.546	0.334	0.125	0.247	0.152
R-Aligned MedSPA	0.569	0.552	0.560	0.341	0.129	0.260	0.169

Table 12: Supplementary comparison of MedSPA with strong general-purpose MLLMs on the MIMIC-CXR test set. General-purpose MLLMs are evaluated without MIMIC-CXR-specific adaptation.

GPT-4V, and Gemini 2.5 Pro. These models are evaluated without MIMIC-CXR-specific adaptation and are therefore included as supplementary references rather than directly controlled baselines.

As shown in Table 12, R-Aligned MedSPA consistently outperforms the evaluated general-purpose MLLMs. Compared with the best-performing general-purpose baseline for each metric, MedSPA improves precision, recall, and CE-F1 by 0.216, 0.073, and 0.161, respectively. The advantage is also observed across the NLG metrics, with improvements of 0.003 in BLEU-1, 0.068 in BLEU-4, 0.020 in ROUGE-L, and 0.032 in METEOR. These results further highlight the strong performance of MedSPA compared with general-purpose MLLMs on MIMIC-CXR.

F.2 LLM-as-a-Judge Evaluation Protocol

We use two independent LLM judges, Gemini 2.5 Pro and GPT-4o, to evaluate the reasoning chains generated by the PR Agent and the final reports generated by the SUM Agent. For both evaluations, the corresponding reference report is provided as the clinical ground truth. Each output receives a single score from 1 to 5 based on its clinical correctness and completeness.

Reasoning Chain Evaluation. The following prompt is used to evaluate whether the generated reasoning chain correctly and sufficiently captures the clinically relevant information in the reference report.

Reasoning Chain Evaluation Prompt

You are a strict radiology evaluator with clinical expertise. Evaluate the generated reasoning chain against the reference radiology report. Assess whether the reasoning is clinically correct and sufficiently covers the clinically relevant findings described in the reference.

Reference report: {Reference Report}

Generated reasoning chain: {Reasoning Chain}

Evaluation criteria:

- Clinical Correctness:** Determine whether the findings in the reasoning chain are consistent with the



GT Report: Enteric and ET tubes are no longer visualized. Degree of pulmonary edema perhaps minimally less extensive when compared to prior. Cardiomegaly is again seen. Retrocardiac region is not well-visualized potentially in part due to overlying soft tissues and atelectasis although underlying infection cannot be excluded."

Reasoning Step

1. Action: Describing initial radiographic context

Reason: This is a semi-erect portable chest radiograph. No endotracheal tube, chest tube, or other foreign bodies are visible. The image shows a standard anterior-posterior projection with the patient in a semi-erect position, which may affect the visualization of dependent lung fields and pleural spaces.

Decision: continue

2. Title: Assessing global lung aeration and volume

Reason: The lung fields appear globally **reduced in volume**, with **the diaphragms elevated and the lung bases appearing compressed**. This is consistent with **low lung volumes**, which may be due to mechanical ventilation, neuromuscular weakness, or restrictive lung disease. The lung parenchyma shows no evidence of hyperinflation or pneumothorax.

Decision: continue

3. Title: Evaluating cardiac silhouette and mediastinum

Reason: The cardiac silhouette is enlarged, with the heart border extending beyond the normal limits of the thoracic cavity. The mediastinum appears centrally located without deviation. The enlargement is consistent with cardiomegaly, which may be contributing to the observed pulmonary venous congestion.

Decision: continue

4. Title: Reviewing pulmonary vascular markings

Reason: The pulmonary vascular markings are prominent and diffuse, with increased opacity and blurring of the vascular margins, particularly in the lower lung zones. This is consistent with pulmonary vascular congestion, likely due to increased pulmonary venous pressure secondary to cardiac dysfunction.

Decision: continue

5. Title: Examining osseous structures and chronic joint changes

Reason: There is a suggestion of **small pleural effusions**, particularly at the costophrenic angles, which **appear blunted and have a meniscus-like contour**. However, the fluid levels are not clearly defined, and the image quality limits definitive confirmation. The effusions are not large or obviously symptomatic.

Decision: summary

Predicted Report: Lung volumes are low. There is cardiomegaly and pulmonary vascular congestion, consistent with heart failure. There is a probable small pleural effusion.

Figure 6: A failure case of MedSPA on MIMIC-CXR. Incorrect or unsupported findings are highlighted in bold.

reference report. Penalize hallucinated findings, contradictions, incorrect anatomical locations or laterality, and unsupported clinical claims.

2. **Completeness:** Determine whether the reasoning chain captures the clinically salient findings in the reference report. Penalize missing important abnormalities, relevant negative findings, or other clinically significant information.

Scoring rubric:

- 5 – Excellent: Clinically correct and complete, with no meaningful errors or omissions.
- 4 – Good: Mostly correct and complete, with only minor errors or omissions that do not substantially affect the clinical interpretation.
- 3 – Moderate: Generally reasonable but contains noticeable errors or misses clinically relevant information.
- 2 – Poor: Contains major clinical errors, unsupported findings, or substantial omissions.
- 1 – Very Poor: Largely incorrect, contradictory, or clinically unreliable.

Output rule: Output exactly one integer from 1 to 5 and nothing else.

Generated Report Evaluation. The final reports are evaluated separately using the same clinical

correctness and completeness criteria.

Generated Report Evaluation Prompt

You are a strict radiology evaluator with clinical expertise. Evaluate the generated radiology report against the reference radiology report. Assess whether the generated report is clinically correct and sufficiently covers the clinically relevant information in the reference.

Reference report: {Reference Report}

Generated report: {Generated Report}

Evaluation criteria:

1. **Clinical Correctness:** Determine whether the findings in the generated report are consistent with the reference report. Penalize hallucinated findings, contradictions, incorrect anatomical locations or laterality, and unsupported clinical claims.
2. **Completeness:** Determine whether the generated report preserves the clinically salient findings in the reference report. Penalize missing important abnormalities, relevant negative findings, or other clinically significant information.

Scoring rubric:

- 5 – Excellent: Clinically correct and complete, with no meaningful errors or omissions.
- 4 – Good: Mostly correct and complete, with only minor errors or omissions that do not substantially affect the clinical interpretation.

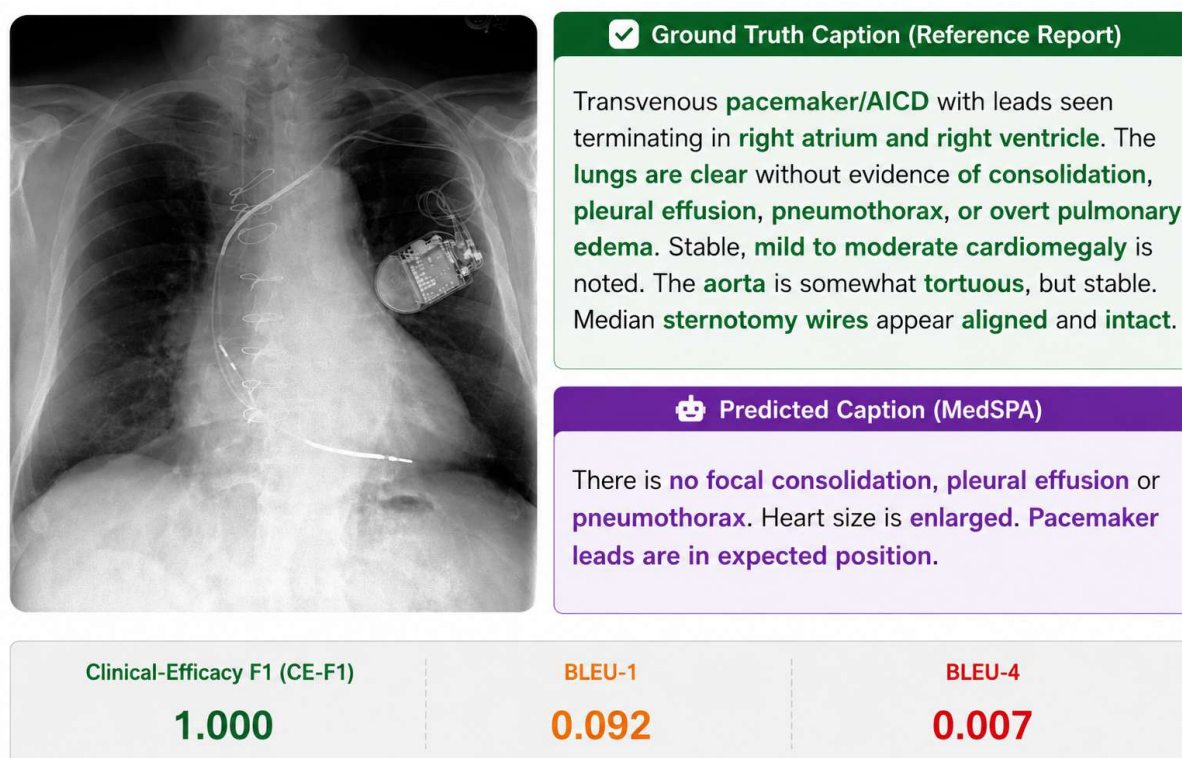


Figure 7: Qualitative example illustrating the discrepancy between clinical-efficacy and lexical-overlap metrics.

- 3 – Moderate: Generally reasonable but contains noticeable errors or misses clinically relevant information.
- 2 – Poor: Contains major clinical errors, unsupported findings, or substantial omissions.
- 1 – Very Poor: Largely incorrect, contradictory, or clinically unreliable.

Output rule: Output exactly one integer from 1 to 5 and nothing else.

G Failure Case Analysis

Figure 6 presents a failure case in which the PR Agent follows a clinically plausible but partially incorrect reasoning trajectory. In Step 2, the model interprets the elevated diaphragms and compressed lung bases as evidence of low lung volumes, although this finding is not described in the reference report. Similarly, in Step 5, the model interprets the apparently blunted costophrenic angles and meniscus-like contour as a possible small pleural effusion, which is also unsupported by the reference.

These steps gradually shift the reasoning toward low lung volume and pleural-effusion findings, while the retrocardiac region is not sufficiently examined. Consequently, the final report preserves

the unsupported low lung volumes and probable small pleural effusion, while missing the atelectatic change and possible underlying infection described in the reference. This example illustrates how locally plausible observations during intermediate reasoning can steer subsequent evidence acquisition toward a suboptimal direction and ultimately propagate into the final report.

H Qualitative Analysis of Clinical and Lexical Alignment

To further investigate the discrepancy between clinical-efficacy and lexical-overlap metrics, we present a representative example from the MIMIC-CXR test set (Figure 7). Despite generating a substantially shorter report with different wording and sentence structure, MedSPA preserves the major clinical findings in the reference report. Specifically, cardiomegaly is correctly identified, while consolidation, pleural effusion, and pneumothorax are correctly reported as absent. Consequently, this example achieves a CE-F1 of 1.0. In contrast, lexical-overlap metrics are very low, with BLEU-1 = 0.092 and BLEU-4 = 0.007. For example, the reference expression “stable, mild to moderate cardiomegaly” is summarized as “heart size is enlarged,” while the detailed description of the

pacemaker leads is compressed into “pacemaker leads are in expected position.” This example illustrates that low lexical overlap does not necessarily imply disagreement in clinically relevant findings.