Learning When to Collaborate: Selective Multi-Agent Medical Reasoning via Uncertainty-Aware Routing

Jiacheng Hou[1], Xueliang Cui[2,4], Dan Lu[5], Ruxin Wang[3,4]*

[1]SIGS, Tsinghua University [2]Southern University of Science and Technology
[3]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[4]Shenzhen University of Advanced Technology, China
[5]West China Hospital, Sichuan University, China.

https://github.com/JayceonHo/SMART-Med

Abstract

Multi-agent systems have emerged as a promising paradigm for multimodal medical reasoning, but existing methods rely on two assumptions that hinder practical deployment: (i) always-active dense collaboration among multiple agents, and (ii) dependence on large cloud-based models that raise privacy and efficiency concerns. In this paper, we challenge both assumptions and propose ***SMART-Med***, a fully local multi-agent method that treats collaboration as a learnable decision rather than a default behavior. *Our key insight is that not every medical query requires costly multi-agent deliberation: many can be reliably solved by a single well-trained agent, while only uncertain cases benefit from expert collaboration.* Based on this insight, ***SMART-Med*** introduces an uncertainty-driven routing mechanism that adaptively decides when to answer directly and when to recruit a small subset of expert agents, effectively converting dense multi-agent reasoning into selective, query-adaptive coordination. To make this routing reliable, we curate a new multimodal medical dataset ***MedVR-3K*** with high-quality reasoning traces and a difficulty hierarchy, and then fine-tune local agents using GRPO with two complementary rewards: a certainty-aware accuracy reward and a vision-language reward model for correctness and reasoning quality respectively. Experiments on four medical visual question answering benchmarks show that ***SMART-Med*** achieves competitive performance compared to existing MAS approaches while substantially reducing inference cost (reducing inference time by more than 3.8× and token usage by 72%). Moreover, after fine-tuning on ***MedVR-3K***, ***SMART-Med*** yields an average improvement of 6.7% across Slake, PathVQA, VQA-RAD, and PMC-VQA. Our results suggest that selective collaboration can be an efficient yet effective alternative to dense collaboration and enable practical, privacy-preserving deployment of multi-agent medical AI.

1 Introduction

Multi-Agent Systems (MAS) have recently emerged as a promising paradigm for complex reasoning in various tasks [2, 4]. In such systems, multiple agents, typically instantiated by Multimodal Large Language Models (MLLMs), are assigned specialized roles and tasks, interacting through structured communication protocols to collaboratively solve complex tasks. Recently, this paradigm has demonstrated promising performance in Medical Visual Question Answering (MedVQA) and broader multimodal medical reasoning [10, 24, 30]. Despite these advances, two critical limitations hinder the practical deployment of MAS in real-world clinical settings.

*Corresponding author. Mails to rx.wang@siat.ac.cn and houjc24@mails.tsinghua.edu.cn.

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

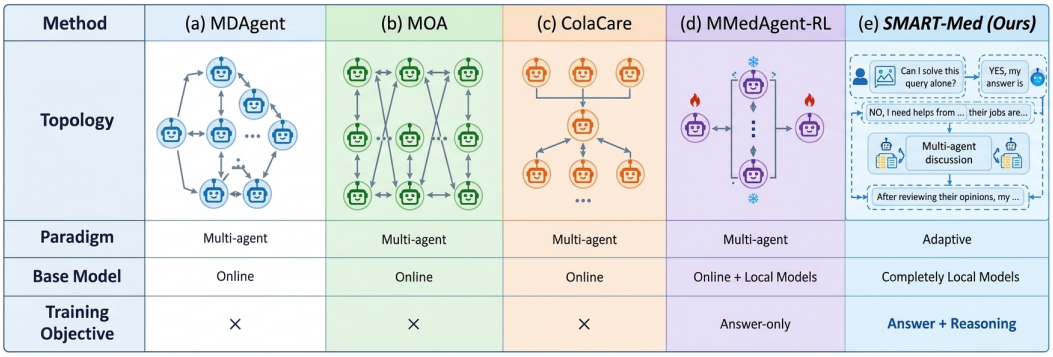


Figure 1: Comparison between existing MAS and our proposed ***SMART-Med***. (a)-(e) respectively depict the collaboration topology, paradigm, base model and training objective of different approaches.

First, existing MAS predominantly rely on large-scale, cloud-based MLLMs (e.g., GPT [11], Gemini [26]), raising concerns regarding data privacy, security, and regulatory compliance. Medical data, particularly multimodal patient records such as MRI, CT, and X-ray scans, are sensitive, and transmitting such data to external servers is often restricted or prohibited in clinical practice. This fundamentally limits the applicability of cloud-dependent MAS solutions. Second, as compared in Fig. 1 (a)-(d), current MAS typically adopt dense communication topologies, where multiple agents engage in multi-round interactions for every query. While this aims to complementarily enhance reasoning quality, they introduce substantial computational overhead and inference latency, making them impractical for resource-constrained clinical environments. Recent studies [35] show that dense collaboration often yields fairly marginal gains on MedQA and MedVQA benchmarks, questioning the necessity of full-agent participation for every query.

To address these challenges, we propose ***SMART-Med***, which operates entirely by local, small-scale MLLMs and incorporates an adaptive collaboration mechanism. Our key insight is that multi-agent collaboration should be *selective rather than universal*: a certain number of queries that can be reliably solved by a single agent, while only a subset of complex cases require coordinated reasoning.

Based on this insight, ***SMART-Med*** introduces two complementary technical components. First, as depicted in Fig. 1 (e), we develop an uncertainty-driven routing mechanism that prevents multi-agent collaboration when the primary agent is certain of its response. In particular, queries with high confidence are answered directly; only uncertain cases trigger dynamic selection of a small subset of expert agents for targeted reasoning. This design significantly reduces inference cost and latency while preserving the benefits of collaboration when necessary.

Second, all agents in ***SMART-Med*** are initialized with local small-scale MLLMs, with fine-tuning applied to improve domain-specific capability. To achieve this, we first construct ***MedVR-3K*** based on existing MedVQA datasets, while ***MedVR-3K*** additionally offers high-quality reasoning chains distilled from advanced MLLMs with multi-step filtering, and an independent judge MLLM is used to filter reasoning traces and assign difficulty scores, yielding a *hierarchical and quality-controlled reasoning-centric dataset*.

Building on this, we fine-tune the local agent using Group Relative Policy Optimization (GRPO) with two tailored rewards. For answer correctness, we introduce a *certainty-aware accuracy reward* that explicitly couples prediction confidence with correctness, encouraging the model to be confident when correct and uncertain otherwise. For reasoning quality, we employ a vision-language reward model, trained on ***MedVR-3K***, to provide fine-grained feedback on the generated reasoning process. Together, these two components improve both reasoning capability and uncertainty calibration, which are critical for enabling reliable selective collaboration. Our contributions can be summarized:

- We propose ***SMART-Med***, the first medical multi-agent system for MedVQA that operates entirely with local, small-scale MLLMs via an uncertainty-driven selective collaboration mechanism, dynamically triggering multi-agent reasoning only for uncertain cases. This design ensures data privacy, significantly reduces inference cost and latency, while preserving collaborative benefits.

- We construct ***MedVR-3K***, a hierarchical reasoning-centric dataset. With the aid of ***MedVR-3K***, we fine-tune the local model using GRPO with two complementary new rewards: a certainty-aware accuracy reward that couples confidence with correctness, and a vision-language reward model that provides fine-grained reasoning quality feedback.
- Extensive experiments on four public MedVQA benchmarks (i.e., Slake, PathVQA, VQA-RAD, and PMC-VQA) show that ***SMART-Med*** achieves competitive or superior performance compared to existing multi-agent frameworks, while significantly reducing inference cost.

2 Related Work

Medical Visual Question Answering. MedVQA aims to automatically answer clinically relevant questions about medical images, requiring precise integration of visual perception and domain knowledge. Conventional MedVQA methods relied on task-specific pipelines. Chen et al. combined CNN-based visual encoding with question classification and retrieval-based answer selection to address long-tailed distributions [3]. Later on, attention-based fusion and consistency constraints are leveraged for improved cross-modal interaction [9, 25]. Recent advances shifted to Transformer-based architectures and vision-language pretraining, leveraging cross-modal attention, contrastive learning [16, 17, 33], and masked multimodal objectives [15] to enhance representation alignment. More recently, foundation models such as LLaVA-Med [13] and Med-Flamingo [21] enable open-ended and few-shot medical reasoning via instruction tuning and generative modeling, yet still suffer from hallucination and limited clinical reliability, motivating extensions such as retrieval-augmented generation (MMed-RAG [31]) and structured reasoning frameworks.

Multi-Agent Collaboration in Medical Analysis. To address the limitations of single-model reasoning, multi-agent systems introduce collaborative inference through role-based interaction and consensus mechanisms. Early approaches rely on predefined discussion protocols or voting strategies, improving robustness but lacking adaptability and efficiency in complex clinical scenarios [6, 24, 28]. Subsequent work explores more flexible coordination strategies. SMOA [14] proposes a sparser collaboration structure by ignoring the results of some agents, producing even better performance. MDAgents [10] dynamically determine the collaboration structures of agents based on task complexity. Despite these advances, most existing methods still depend on heuristic interaction designs or incur substantial coordination overhead, motivating recent learning-based approaches such as MMedAgent-RL [30], which optimize agent collaboration and information fusion through reinforcement learning for more scalable and reliable medical reasoning.

3 Methodology

3.1 Problem Formulation

In MedVQA, the input consists of a medical image $x \in \mathcal{X}$ (e.g., X-ray, CT, MRI, pathology slides) and a natural language question $q \in \mathcal{Q}$. The goal is to generate an answer $y \in \mathcal{Y}$ accompanied by a reasoning chain r. In a multi-agent system, we define a set of K agents $\mathcal{A} = \{A_1, \ldots, A_K\}$, where each agent is instantiated by a MLLM with parameters θ_k. Agents may specialize in different reasoning styles or roles (e.g., visual grounding, clinical knowledge, verification). Current MAS generally produce answers following a dense collaboration protocol:

$$y = \mathcal{F}_{\text{dense}}(x, q; \mathcal{A}), \tag{1}$$

where multiple agents engage in multi-round communication. In contrast, we aim to learn a routing policy π_θ that smartly switches between single-agent and multi-agent modes:

$$y = \mathcal{F}_{\text{sparse}}(x, q; \mathcal{A}, \pi_\theta), \tag{2}$$

where π_θ determines whether multi-agent collaboration is needed and, if so, how many agents should be activated, conditioned on the input. This transforms the dense communication topology of a MAS into a sparse, query-adaptive structure by activating only a few (or even a single) agent.

3.2 *SMART-Med*: Learning When to Collaborate

In real-world clinical practice, the first step when a patient visits a hospital is triage, assigning the patient to an appropriate department based on symptoms. In ***SMART-Med***, we employ a router agent

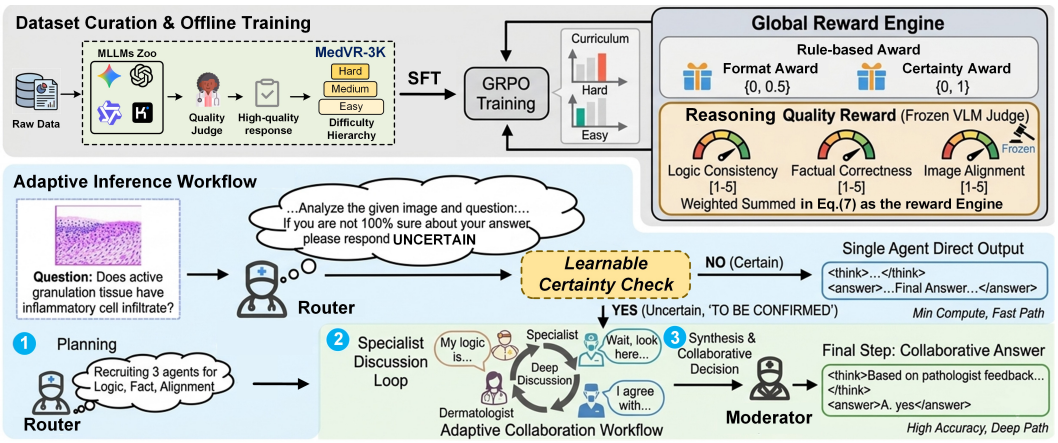


Figure 2: Overview of ***SMART-Med***. The framework unifies reasoning-centric training and uncertainty-aware routing in a closed loop: ***MedVR-3K*** and reward-guided optimization improve both reasoning capability and uncertainty calibration, while the learned router adaptively decides whether to answer directly or trigger selective expert collaboration.

that plays the role of a general practitioner. Its parameters are our optimization target π_θ, which decides when to invoke multi-agent collaboration.

Concretely, we instantiate a local MLLM as the router. It generates an initial assessment $\tilde{y}$ and appends a certainty token after $\tilde{y}$ for each medical query:

$$\tilde{y} \in \mathcal{Y} \cup \{\texttt{CERTAIN} \mid \texttt{UNCERTAIN}\}, \tag{3}$$

where the certainty token indicates whether the router is sufficiently certain about its initial assessment $\tilde{y}$. As shown in Fig. 2, if the router is confident (outputs `CERTAIN`), ***SMART-Med*** directly adopts $\tilde{y}$ as the final answer and skips the expensive multi-agent collaboration. Otherwise, the router recruits a small subset of expert agents to perform collaborative reasoning. In this case, a moderator agent gathers the analysis reports from the recruited experts and makes the final decision. This discussion-and-decision workflow follows prior work [24, 30], with the designed prompts and role descriptions provided in Appendix F. Thus, the workflow of ***SMART-Med*** is fully adaptive, operating in either single-agent or multi-agent mode depending on the given query.

Reward-Shaped Optimization. We employ Group Relative Policy Optimization (GRPO) [23] to jointly optimize answer correctness, uncertainty calibration, reasoning quality, and output format compliance. First, we define a format reward:

$$R_{\text{FMT}}(o_i) = \mathbb{I}[o_i \text{ strictly matches the required output schema}], \tag{4}$$

which enforces structural constraints (e.g., certainty level, reasoning chain, final answer) on the generated response o_i.

Next, we introduce two customized rewards: a *Certainty-aware Accuracy Reward (CAR)* and a *Reasoning-quality Reward (RR)*. Let y_i be the answer extracted from o_i, c_i the certainty decision ($c_i = 1$ for `CERTAIN`, 0 for `UNCERTAIN`), and $y^\star$ the ground-truth answer. We define CAR as:

$$R_{\text{CAR}}(o_i) = \mathbb{I}(y_i = y^\star) \cdot \big[c_i + (1 - c_i)\alpha\big] - \mathbb{I}(y_i \neq y^\star) \cdot c_i\beta, \tag{5}$$

where $\alpha \in (0, 1)$ gives reduced credit to correct but uncertain outputs, and $\beta > 0$ penalizes confident mistakes. This reward induces a selective prediction objective with a *reject option*, encouraging the model to output `UNCERTAIN` when its direct answer is likely wrong. For RR, we use a frozen Vision-Language Model (VLM) fine-tuned on our dataset:

$$R_{\text{RR}}(o_i) = f_{\text{VLM}}(x, q, r_i, y_i), \tag{6}$$

where r_i is the generated reasoning trace. The combined raw reward is:

$$\bar{R}_i = \lambda_{\text{fmt}} R_{\text{FMT}}(o_i) + \lambda_{\text{car}} R_{\text{CAR}}(o_i) + \lambda_{\text{rr}} R_{\text{RR}}(o_i), \tag{7}$$

with coefficients $\lambda_{\text{fmt}}, \lambda_{\text{car}}, \lambda_{\text{rr}}$ controlling the relative importance of format compliance, answer correctness/certainty, and reasoning quality.

Following standard GRPO, we sample a group of G candidate outputs $\{o_i\}_{i=1}^{G}$ for each input (x, q) and standardize the rewards within the group:

$$A_i = \frac{\bar{R}_i - \mu_R}{\sigma_R}, \tag{8}$$

where μ_R and σ_R are the group mean and variance. The standardized advantage A_i measures how much better a response is relative to its group under our rewards. For simplicity, we omit the KL constraint, and the objective can be expressed in an approximate form as follows:

$$-\mathbb{E}_{(x,q)\sim\mathcal{D},\,\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\Big(r_{i,t}(\theta)A_{i,t},\,\text{clip}\left(r_{i,t}(\theta),1-\epsilon,1+\epsilon\right)A_{i,t}\Big)\right], \tag{9}$$

where π_{ref} is a frozen reference policy, and β_{KL} controls the deviation from the reference. This objective jointly encourages correct answers, calibrated uncertainty, high-quality reasoning traces, and strict format adherence.

Theoretical Justification. We provide a decision-theoretic interpretation of our certainty-aware design, together with its implications for optimization under GRPO. Let $z \in \{0, 1\}$ denote whether the router's direct answer is correct, and let

$$p(x, q) = \Pr(z = 1 \mid x, q) \tag{10}$$

denote the conditional correctness probability of the direct answer. In the standard setting without deferral, the model always produces an answer, yielding the expected risk under 0-1 loss:

$$R_{\text{no_rej}}(x, q) = 1 - p(x, q). \tag{11}$$

Our formulation introduces an additional decision: the router may either answer directly or declare `UNCERTAIN`, which triggers multi-agent collaboration. This can be interpreted as a form of *deferred decision-making*, where uncertain cases are routed to a collaborative reasoning process. To characterize the reward-induced routing preference, consider an idealized decision setting in which $p(x, q)$ is calibrated and the router selects the action maximizing its expected certainty-aware reward. Under the certainty-aware reward in Eq. (5), the expected utilities of the two actions can be written as

$$U_{\text{ans}} = p(x, q) - (1 - p(x, q))\beta, \qquad U_{\text{def}} = p(x, q)\alpha, \tag{12}$$

where $\alpha \in (0, 1)$ gives partial credit to correct but deferred outputs, and $\beta > 0$ penalizes overconfident errors. Comparing the two utilities gives the reward-induced Bayes-optimal threshold:

$$U_{\text{ans}} \geq U_{\text{def}} \quad \Longleftrightarrow \quad p(x, q) \geq \tau_{\text{Bayes}}, \qquad \tau_{\text{Bayes}} = \frac{\beta}{1 - \alpha + \beta}. \tag{13}$$

Thus, the certainty-aware reward induces a non-trivial utility boundary: direct answering is preferred when the estimated correctness probability exceeds τ_{Bayes}, whereas deferral is preferred otherwise. The detailed derivation of this reward-induced threshold is provided in Appendix B. Importantly, τ_{Bayes} characterizes the *reward-induced theoretical threshold* rather than the exact decision boundary of the GRPO-trained router. In practice, GRPO learns a parameterized policy $\pi_\theta(a \mid x, q)$, and the optimized router is given by

$$\theta^\star \approx \arg\max_{\theta} J_{\text{GRPO}}(\theta; \alpha, \beta), \tag{14}$$

with the actual routing decision determined by

$$a_{\theta^\star}(x, q) = \arg\max_{a\in\{\text{ANS},\text{DEF}\}} \pi_{\theta^\star}(a \mid x, q). \tag{15}$$

Accordingly, the actual learned routing boundary is implicitly defined by

$$\pi_{\theta^\star}(\text{ANS} \mid x, q) = \pi_{\theta^\star}(\text{DEF} \mid x, q). \tag{16}$$

The derivation above provides a principled decision-theoretic interpretation of how α and β shape the relative utility of direct answering and deferral. In practice, the resulting routing behavior is learned

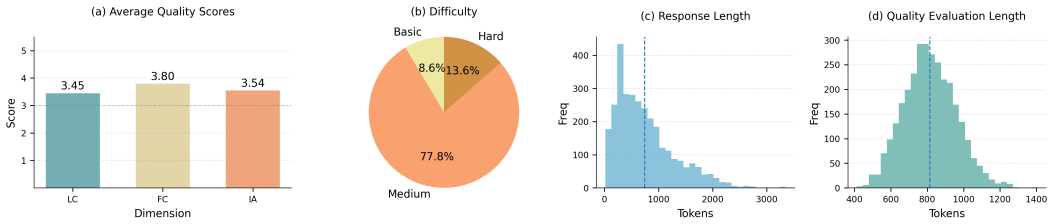


Figure 3: Statistics of our constructed ***MedVR-3K***. (a)-(d) respectively present the average quality score, distribution of difficulty level, response length and quality assessment length.

implicitly through GRPO and depends not only on α and β, but also on the training data, optimization dynamics, and the learned policy.

From an optimization perspective, a standard reward $R_{\text{base}} = \mathbb{I}(y = y^\star) + \lambda_{\text{fmt}}\mathbb{I}(\text{format})$ cannot distinguish overconfident errors from calibrated deferral, since both are treated identically with respect to answer correctness, which optimizes in a more narrow landscape and cannot penalize over-confident or over-conservative cases.

3.3 *MedVR-3K*: Medical Reasoning-Centric Dataset

To improve the reasoning capability of local MLLMs, we construct ***MedVR-3K***, a reasoning-centric multimodal medical dataset. As shown in Fig. 3, besides the medical question and image (e.g., MRI, CT, pathology), each sample in ***MedVR-3K*** additionally comprises (i) a gold response with a step-by-step reasoning chain, (ii) a difficulty label, and (iii) a quality score along with a justification. Unlike prior MAS that only supervise final answers [30], ***MedVR-3K*** explicitly provides the reasoning process, offering a more informative supervision signal for multimodal reasoning. Examples are shown in Appendix E.

Data Collection. As illustrated in Fig. 2, we collect question-image pairs from the training splits of four public benchmarks: Slake [19], PMC-VQA [34], VQA-RAD [12], and PathVQA [7]. For each sample (x, q), we query a set of strong candidate MLLMs (i.e., GPT-5.4 [22], Gemini 3 Pro [26], Qwen3-Max [1], and Kimi K2.5 [27]) and require each response to include a structured, step-by-step reasoning chain together with a final answer. Among the candidate responses, we retain only those whose final answers match the ground-truth label. This first-stage filtering ensures that the dataset contains reliable reasoning traces instead of fluent but incorrect explanations.

Annotation and Quality Control. After correctness filtering, we perform a second-stage quality control using GPT-5.4 as an independent judge J. Instead of a single scalar score, the judge evaluates each response along three dimensions: *Logic Consistency*, *Factual Correctness*, and *Image Alignment*. Formally, for a candidate response r, the judge produces a score vector

$$\mathbf{s} = \big(s_{\text{logic}},\, s_{\text{fact}},\, s_{\text{img}}\big) = J(x, q, r), \tag{17}$$

where the score s is in range of $[1, 5]$. We adopt a dimension-wise filtering strategy and retain a response only if it satisfies all quality thresholds (denoted by γ):

$$s_{\text{logic}} \geq \gamma_{\text{logic}}, \quad s_{\text{fact}} \geq \gamma_{\text{fact}}, \quad s_{\text{img}} \geq \gamma_{\text{img}}. \tag{18}$$

This multi-criteria filtering ensures that retained reasoning traces are not only correct in final answers but also logically coherent, factually accurate, and well-grounded in visual evidence. Compared to single-score filtering, this approach avoids cases where strong performance in one dimension compensates for deficiencies in others, yielding more reliable supervision. Subsequently, as shown in Fig. 3, the average quality score of ***MedVR-3K*** reaches around 3.60.

Difficulty Annotation. Beyond quality control, we further leverage the judge model to annotate the difficulty of each sample. Specifically, the judge assigns a difficulty label d based on both question complexity and visual complexity:

$$d = J(x, q, r) \in \{\texttt{basic}, \texttt{medium}, \texttt{hard}\}. \tag{19}$$

Table 1: Accuracy comparison with state-of-the-art baseline models on Slake, PathVQA, VQA-RAD and PMC-VQA datasets under multi-choice and free-form questions.

Category	Methods	Multiple-Choice Questions (MC)				Free-form Questions (FF)		
		Slake	PathVQA	VQA-RAD	PMC-VQA	Slake	PathVQA	VQA-RAD
Single agent (Specialized model)	BiomedGPT	53.5±0.3	57.0±0.7	52.0±0.2	41.5±0.7	54.7±0.8	39.5±2.2	54.8±3.4
	Med-Flamingo	61.5±1.5	58.5±3.0	54.5±1.9	38.0±2.0	57.9±1.1	34.7±2.8	51.5±4.2
	LLaVA-Med	59.0±0.1	52.5±0.1	60.5±0.4	36.0±1.3	59.6±0.4	48.5±0.5	65.8±1.7
	HuatuoGPT-Vision	69.5±1.8	59.0±1.4	64.0±1.6	41.0±2.2	59.6±1.8	45.9±1.6	63.8±0.6
Single agent (Local models)	LLaVA-v1.6-7B	65.0±0.5	52.0±1.2	54.0±0.7	32.0±2.1	52.4±2.8	31.3±0.7	50.8±2.9
	Minstral-3-3B-Instruct	68.0±1.2	51.0±0.4	55.5±1.1	45.0±2.3	67.2±1.2	33.3±1.9	45.7±0.3
	Minstral-3-8B-Instruct	70.0±0.8	63.0±1.7	57.0±1.5	38.5±1.1	77.3±1.0	41.7±0.3	62.9±0.7
	Qwen3-VL-2B-Instruct	76.0±0.5	58.5±1.1	66.7±1.0	30.2±1.3	73.6±2.2	40.5±1.6	58.0±0.1
	Qwen3-VL-4B-Instruct	76.1±1.4	64.6±2.1	70.8±1.8	35.3±2.8	75.4±0.8	44.3±2.5	62.6±4.2
	Qwen3-VL-8B-Instruct	71.9±2.3	59.0±0.4	64.2±0.5	50.2±2.4	73.3±0.7	45.8±3.0	60.3±5.0
	Qwen3-VL-8B-Thinking	73.6±0.8	59.0±1.0	60.5±2.3	33.8±1.0	67.6±2.4	49.6±0.8	53.6±1.8
Multi-agent (Local models)	MedAgents	72.0±1.4	73.7±2.2	72.0±0.1	52.5±0.1	74.0±0.8	48.8±1.6	67.0±3.0
	MDAgents	70.2±1.2	75.6±1.0	68.6±2.0	50.6±1.0	69.1±1.9	46.6±0.6	57.4±4.3
	MOA	79.0±1.0	69.2±2.1	71.5±0.5	51.5±1.1	72.9±0.7	52.6±4.0	62.3±2.7
	Reconcile	72.8±1.0	64.2±1.2	66.0±0.5	49.0±1.0	58.1±1.0	43.1±1.8	57.5±4.3
	SMOA	68.0±1.1	62.2±3.7	63.6±1.2	49.3±2.6	60.3±1.0	38.9±1.1	63.0±1.5
	ColaCare	71.3±1.5	60.8±2.0	65.3±0.6	50.0±0.8	60.5±1.8	45.2±1.0	53.8±2.6
Adaptive agent (Local models)	***SMART-Med*** (Ours)	76.7±1.0	68.6±1.3	72.5±0.5	48.5±2.1	78.2±1.3	48.7±0.2	64.1±3.2
	SMART-Med * (Ours)	**87.0**±0.5	**77.0**±1.1	**74.0**±1.6	**72.5**±1.0	**84.3**±2.5	**58.5**±3.5	**69.4**±3.7

Note: MC: Multiple-choice question answering; FF: Free-form (open-ended) question answering. **Bold** indicates the best performance, and Underlined indicates the second-best performance per column (dataset and task). All scores are reported as mean±standard deviation by testing 5 times independently. For fair comparison, we use Qwen3-VL-4B-Instruct as the base model for MAS approaches. ***SMART-Med*** * denotes our method after fine-tuning on ***MedVR-3K***.

The difficulty is determined by factors such as the number of reasoning steps, the level of domain knowledge required, and the ambiguity or complexity of visual evidence. In particular, `basic` samples typically involve direct visual recognition, while `medium` and `hard` samples require multi-step reasoning, cross-region inference, or integration of domain-specific knowledge. This hierarchical annotation enables curriculum learning, allowing the model to progressively learn from simple to complex reasoning patterns. Detailed prompts are provided in Appendix F.

After the two-stage filtering and hierarchical annotation, ***MedVR-3K*** contains 3,477 multimodal medical QA samples with structured reasoning chains, which serves as a hierarchical MedVQA dataset specifically designed for reasoning-centric multimodal medical learning.

4 Experiments

4.1 Setup

We evaluate our method on four widely-used public MedVQA datasets: Slake [19], PathVQA [7], VQA-RAD [12], PMC-VQA [34], which cover diverse imaging modalities and clinical scenarios. We compare our method with a series of single agent and multi-agent methods as follows:

- **Single Agent.** We adopt different types and sizes of representative open-source MLLMs to initialize a single agent with basic prompts to serve as a reference, including Qwen3-VL series [1], LLaVA [20], Ministral 3 series [18], and the specialized medical models, the BiomedGPT [32], MedFlamingo [21], HuatuoGPT-Vision [5], and LLaVA-Med [13].
- **Multi-agent Collaboration.** We also implement the latest and state-of-the-art MAS methods: MDAgents [10], MedAgent [24], MOA [28], SMOA [14], Reconcile [6], and ColaCare [29].

More information about the datasets, implementation details and evaluation metrics can be found in Appendix A.

4.2 Results Analysis

Quantitative Comparison. As shown in Table 1, without any fine-tuning, ***SMART-Med*** already achieves competitive performance among local multi-agent approaches. For example, on FF tasks, ***SMART-Med*** attains 78.2% on Slake, outperforming the best local baseline (77.3% from Minstral-3-8B-Instruct) and surpassing most multi-agent methods. On PathVQA and VQA-RAD, it achieves 48.7% and 64.1%, respectively, remaining competitive with dense multi-agent frameworks such as

Table 2: Accuracy comparison with representative multi-agent collaboration methods after fine-tuning on our curated ***MedVR-3K*** dataset using identical fine-tuning protocol.

Method	MC				FF		
	Slake	PathVQA	VQA-RAD	PMC-VQA	Slake	PathVQA	VQA-RAD
MOA*	85.6±1.8	70.0±1.4	**75.5**±1.1	65.3±0.6	79.0±0.9	56.0±0.4	61.3±0.7
MDAgent*	86.7±0.9	75.3±1.7	71.9±2.1	67.7±1.9	80.9±0.8	50.5±2.2	53.9±5.0
SMART-Med * (Ours)	**87.0**±1.0	**77.0**±1.1	74.0±1.6	**72.5**±1.0	**84.3**±2.5	**58.5**±3.5	**69.4**±3.7

Table 3: Reasoning quality comparison. LC, FC, and IA respectively represent the Logic Consistency, Factual Correctness and Image Alignment.

Score	MDAgent	MedAgent	SMOA	MOA	ColaCare	*SMART-Med*	*SMART-Med* *
LC	2.89	3.31	3.48	3.42	3.28	3.24	**3.55**
FC	2.80	3.11	3.37	3.12	3.20	2.96	**3.38**
IA	2.56	3.19	3.29	3.40	3.18	3.14	**3.57**
Avg.	2.75	3.20	3.38	3.31	3.22	3.11	**3.50**

MOA (52.6%, 62.3%) and MedAgents (48.8%, 67.0%). On MC tasks, ***SMART-Med*** also performs strongly, reaching 76.7% on Slake and achieving the second-best result on VQA-RAD (72.5%), while maintaining reasonable performance on the more challenging PMC-VQA (48.5%).

After fine-tuning on ***MedVR-3K***, ***SMART-Med*** * further achieves substantial improvements across all benchmarks. On MC tasks, ***SMART-Med*** * reaches 87.0% on Slake, 77.0% on PathVQA, 74.0% on VQA-RAD, and 72.5% on PMC-VQA, exceeding the best baselines by +8.0%, +1.4%, +2.0%, and +20.0% points, respectively. On FF tasks, it achieves 84.3% on Slake, 58.5% on PathVQA, and 69.4% on VQA-RAD, establishing new state-of-the-art results among local models. Notably, the largest gains are observed on PMC-VQA and PathVQA, which involve more complex reasoning, highlighting the effectiveness of our reasoning-centric training and reward design. Overall, these results show that while ***SMART-Med*** provides a strong and efficient zero-shot solution, ***SMART-Med*** * further unlocks its potential through reasoning supervision and uncertainty-aware optimization, leading to consistently superior performance.

To disentangle the accuracy gains of ***SMART-Med*** * from the effect of fine-tuning on ***MedVR-3K*** itself, we additionally fine-tune two representative baselines, MOA and MDAgent, with the identical ***MedVR-3K*** corpus and training pipeline (replacing the certainty-aware reward with the standard accuracy reward, since neither models a certainty token). Table 2 shows that ***MedVR-3K*** consistently benefits all methods, yet ***SMART-Med*** * still achieves the best overall performance after controlling for identical post-training, indicating that uncertainty-aware routing provides complementary benefits beyond domain-specific fine-tuning alone.

Qualitative Comparison. We randomly select 100 samples, covering both FF and MC questions for reasoning quality comparison. Similarly, we employ the GPT-5.4 as the judge model, and we evaluate the generated reasoning along three dimensions: Logic Consistency (LC), Factual Correctness (FC), and Image Alignment (IA). The prompts for the judge model can be found in Appendix F. The results are reported in Table 3. It can be seen that ***SMART-Med*** * achieves the best performance across all dimensions, with an average score of 3.50, outperforming prior multi-agent methods such as SMOA (3.38) and MOA (3.31). In particular, ***SMART-Med*** * shows strong improvements in Image Alignment (3.57 vs. 3.40 of MOA) and Logic Consistency (3.55 vs. 3.48 of SMOA), indicating that our method generates reasoning that is better grounded in visual evidence while maintaining coherent logical structure. Comparing ***SMART-Med*** and ***SMART-Med*** *, we observe consistent gains across all metrics (Avg.: 3.11 $\rightarrow$ 3.50), validating the effectiveness of our GRPO training with certainty-aware and reasoning rewards. More validations and analysis can be found in Appendix C.

Efficiency Analysis. We evaluate efficiency of existing MAS approaches in terms of latency, token usage, and communication overhead on 100 randomly-sampled questions. From Fig. 4, it has been shown that ***SMART-Med*** is significantly more efficient than other approaches, requiring less than 10 seconds for every query on average, corresponding to a 3.8$\times$ speedup over the fastest baseline (i.e., MDAgents) and up to 11.9$\times$ over dense collaboration methods (i.e., ColaCare). In addition, ***SMART-Med*** merely consumes around 600 input tokens and 100 output tokens per query,

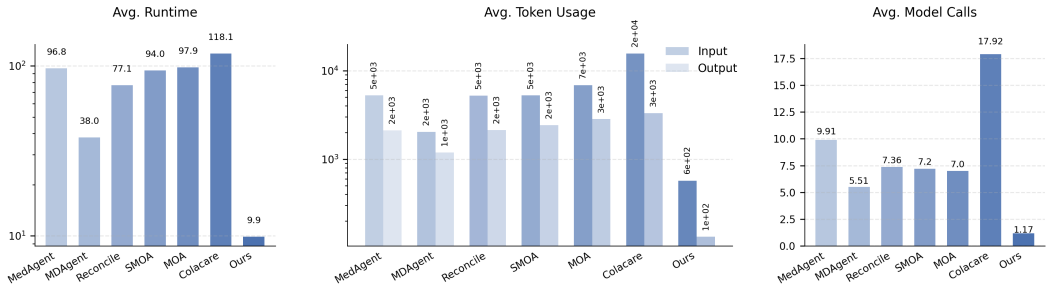


Figure 4: Efficiency comparison of ***SMART-Med*** and other MAS methods. We compare the average runtime (second), average I/O token usage and the average model calls (times) per question.

Table 4: Ablation study results on MC and FF questions.

Model	MC				FF		
	Slake	PathVQA	VQA-RAD	PMC-VQA	Slake	PathVQA	VQA-RAD
Single agent							
	76.1±1.4	64.6±2.1	70.8±1.8	35.3±2.8	75.4±0.8	44.3±2.5	62.6±4.2
Multi-agent							
+ Dense collaboration	78.5±0.3	66.7±2.4	71.3±2.2	50.1±0.5	75.7±2.8	51.2±1.7	64.5±1.7
+ Uncertainty routing	76.7±1.0	68.6±1.3	72.5±0.5	48.5±2.1	78.2±1.3	48.7±0.2	64.1±3.2
+ Certain. reward	85.5±1.0	75.0±0.9	73.1±1.2	69.5±1.1	82.2±1.8	56.8±1.2	68.7±2.6
+ Reason. reward	**87.0**±0.5	**77.0**±1.1	**74.0**±1.6	**72.5**±1.0	**84.3**±2.5	**58.5**±3.5	**69.4**±3.7

reducing communication cost by over 72% and 88%, respectively. It also requires only 1.17 model calls per question, far less than 5.51–17.92 in prior methods. These results verify that selective, uncertainty-aware routing substantially reduces computation and communication overhead.

4.3 Ablation Study

Effect of Agent Collaboration. From Table 4, the dense collaboration improves MC accuracy on PMC-VQA from 35.3% to 50.1% (+14.8%) and PathVQA from 64.6% to 66.7%, while also bringing gains on FF tasks such as PathVQA (44.3% to 51.2%). Replacing dense collaboration with uncertainty-driven routing achieves comparable or even better performance without any additional training. Notably, uncertainty routing outperforms dense collaboration on PathVQA (66.7% to 68.6%) and VQA-RAD (71.3% to 72.5%) in MC, and achieves higher FF accuracy on Slake (75.7% to 78.2%). These results confirm that multi-agent collaboration should be *selective rather than universal*: many queries can be reliably solved by a single agent, while only a subset of complex cases benefit from the agent collaboration.

Effect of Agent Learning. Training with GRPO using our proposed Certainty-Aware Accuracy Reward (CAR) further boosts performance across all benchmark. For instance, MC accuracy on PMC-VQA improves significantly from 48.5% to 69.5%, and PathVQA from 68.6% to 75.0%, while FF performance on Slake increases from 78.2% to 82.2%. Furthermore, incorporating the Reasoning-quality Reward (RR) leads to additional gains, particularly on FF tasks where reasoning quality is critical (e.g., PathVQA: 56.8% to 58.5%, Slake: 82.2% to 84.3%), and also improves MC performance (e.g., PMC-VQA: 69.5% to 72.5%). These results show that the combination of CAR and RR enables the model to both make calibrated routing decisions and generate higher-quality reasoning, resulting in the best overall performance.

5 Conclusion

In this paper, we present ***SMART-Med***, an adaptive and efficient multi-agent framework for medical multimodal reasoning that leverages uncertainty-driven sparse routing to enable selective collaboration. To support effective training, we construct ***MedVR-3K***, a reasoning-centric MedVQA benchmark with hierarchical difficulty, and propose a unified reward design that jointly models correctness, uncertainty, reasoning quality, and structural consistency. Experiments on multiple benchmarks show that ***SMART-Med*** achieves state-of-the-art performance among local models while

significantly reducing computational overhead and preserving data privacy. These results highlight the potential of fully local agent systems with reasoning-centric fine-tuning for building practical, reliable, and deployable medical AI systems.

Acknowledgment

This paper was supported in part by the National Key R&D Program of China under Grant 2022YFA1008300, in part by the National Natural Science Foundation of China under Grant 12471308, and in part by the Excellent Young Scholars of Shenzhen under Grant RCYX20231211090247060.

References

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

[2] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations*.

[3] Guanqi Chen, Haifan Gong, and Guanbin Li. Hcp-mic at vqa-med 2020: Effective visual representation for medical visual question answering. In *CLEF (Working Notes)*, pages 1–8, 2020.

[4] Huaben Chen, Wenkang Ji, Lufeng Xu, and Shiyu Zhao. Multi-agent consensus seeking via large language models. *arXiv preprint arXiv:2310.20151*, 2023.

[5] Junying Chen, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, and Benyou Wang. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv*, 2024.

[6] Justin Chen, Swarnadeep Saha, and Mohit Bansal. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7066–7085, 2024.

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

[8] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representation*, 1(2):3, 2022.

[9] Xiaofei Huang and Hongfang Gong. A dual-attention learning network with word and sentence embedding for medical visual question answering. *IEEE Transactions on Medical Imaging*, 43(2):832–845, 2023.

[10] Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik S Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae W Park. Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems*, 37:79410–79452, 2024.

[11] Diane M. Korngiebel and Sean D. Mooney. Considering the possibilities and pitfalls of generative pre-trained transformer 3 (gpt-3) in healthcare delivery. *npj Digital Medicine*, 4(1):93, 2021.

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

[13] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*, 2023.

[14] Dawei Li, Zhen Tan, Peijia Qian, Yifan Li, Kumar Chaudhary, Lijie Hu, and Jiayi Shen. Smoa: Improving multi-agent large language models with s parse m ixture-o f-a gents. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 54–65. Springer, 2025.

[15] Pengfei Li, Gang Liu, Jinlong He, Zixu Zhao, and Shenjun Zhong. Masked vision and language pre-training with unimodal and multimodal contrastive losses for medical visual question answering. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 374–383. Springer, 2023.

[16] Pengfei Li, Gang Liu, Lin Tan, Jinying Liao, and Shenjun Zhong. Self-supervised vision-language pretraining for medial visual question answering. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2023.

[17] Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 525–536. Springer, 2023.

[18] Alexander H Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, et al. Ministral 3. *arXiv preprint arXiv:2601.08584*, 2026.

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

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

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

[22] OpenAI. Gpt-4 technical report, 2024.

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

[24] Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. Medagents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 599–621, 2024.

[25] Sergio Tascon-Morales, Pablo Márquez-Neila, and Raphael Sznitman. Consistency-preserving visual question answering in medical imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 386–395. Springer, 2022.

[26] Gemini Team. Gemini: A family of highly capable multimodal models, 2024.

[27] Kimi Team, Yifan Bai, Yiping Bao, Y Charles, Cheng Chen, Guanduo Chen, Haiting Chen, Huarong Chen, Jiahao Chen, Ningxin Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

[28] Junlin Wang, WANG Jue, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. In *The Thirteenth International Conference on Learning Representations*, 2024.

[29] Zixiang Wang, Yinghao Zhu, Huiya Zhao, Xiaochen Zheng, Dehao Sui, Tianlong Wang, Wen Tang, Yasha Wang, Ewen Harrison, Chengwei Pan, et al. Colacare: Enhancing electronic health record modeling through large language model-driven multi-agent collaboration. In *Proceedings of the ACM on Web Conference 2025*, pages 2250–2261, 2025.

[30] Peng Xia, Jinglu Wang, Yibo Peng, Kaide Zeng, Zihan Dong, Xian Wu, Xiangru Tang, Hongtu Zhu, Yun Li, Linjun Zhang, Shujie Liu, Yan Lu, and Huaxiu Yao. Mmedagent-rl: Optimizing multi-agent collaboration for multimodal medical reasoning. In *International Conference on Learning Representations*, 2026.

[31] Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. Mmed-rag: Versatile multimodal rag system for medical vision language models. In *International Conference on Learning Representations*, 2025.

[32] Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D Davison, Hui Ren, et al. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, pages 1–13, 2024.

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

[34] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Development of a large-scale medical visual question-answering dataset. *Communications Medicine*, 4:277, 2024.

[35] Yinghao Zhu, Ziyi He, Haoran Hu, Xiaochen Zheng, Xichen Zhang, Zixiang Wang, Junyi Gao, Liantao Ma, and Lequan Yu. Medagentboard: Benchmarking multi-agent collaboration with conventional methods for diverse medical tasks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

Appendix

- §A Implementation Details
  - Training Details
  - Hyper-Parameters
  - Datasets
  - Evaluation Metrics
- §B Derivation of the Reward-Induced Decision Threshold
- §C Additional Analysis
  - Routing Validation
  - Risk–Coverage Analysis
  - False-Certainty Safety Analysis
  - Statistical Significance
  - Extra Judge Evaluation
- §D Examples of ***SMART-Med***
  - Example 1: Pathology Question on Inflammatory Infiltrate Characterization
  - Example 2: Multiple-Choice Question on Muographic Map Interpretation
  - Discussion
- §E Showcases of ***MedVR-3K***
- §F Prompts Design
  - Prompts for Building ***MedVR-3K***
  - Prompts for Building ***SMART-Med***

A Implementation Details

Training Details. We adopt the Qwen3-VL-4B-Instruct [1] as the foundation model due to the trade-off between model scale and performance, as well as its superior instruction-following capability in multimodal settings. We further post-train the model on our constructed ***MedVR-3K*** to enhance its reasoning ability and uncertainty estimation. Specifically, we employ Low-Rank Adaptation (LoRA) [8] to fine-tune all linear layers of the model for both supervised fine-tuning (SFT) and reinforcement learning. During SFT, we train the model for 5 epochs with a learning rate of 1×10^{-4}. For the subsequent GRPO stage, we adopt a smaller learning rate of 1×10^{-6} and fine-tune for 1 epoch to stabilize policy optimization. We follow a curriculum learning strategy during training. The model is first trained on samples with lower difficulty levels and progressively exposed to more challenging samples, enabling it to gradually acquire complex reasoning capabilities. In the GRPO stage, we focus exclusively on high-difficulty samples to further push the upper bound of the model's reasoning performance and decision-making ability under uncertainty. For the VLM reward model, we employ the Qwen3-VL-2B-Instruct and fine-tune it on the response-quality pairs offered by ***MedVR-3K*** for 10 epochs. During GRPO fine-tuning, it will be frozen and used to output the quality rewards. All experiments are conducted on 8 NVIDIA Tesla H100 GPUs. We implement our training pipeline based on the ms-swift framework. We set the number of generations for GRPO is 6 and the number of gradient accumulation steps is 3.

Hyper-Parameters. In our experiments, we set the maximum number of specialists to 3 and discussion rounds to 2 for both baselines and our ***SMART-Med***. All agents are initialized from Qwen3-VL-4B-Instruct, except for ***SMART-Med*** *, which uses a version fine-tuned on our ***MedVR-3K***. Unless otherwise noted, our implementation follows the official repository.

Table 5: Dataset statistics. We compare the modality, size, question type of 4 datasets as well as our constructed ***MedVR-3K*** that are used in our work.

	Slake	PathVQA	VQA-RAD	PMC-VQA	***MedVR-3K***
Size	4,717	26,516	1,778	4,000	3,477
Answer	✓	✓	✓	✓	✓
Reasoning	✗	✗	✗	✗	✓
Difficulty	✗	✗	✗	✗	✓
Modality	Radiology images	Pathology slides	Radiology images	Radiology images	Pathology + Radiology
Question Type	YES/NO/OPEN	YES/NO/OPEN	YES/NO/OPEN	A/B/C/D	YES/NO/A/B/C/D

Datasets. In this work, we employ 4 popular MedVQA datasets as follows:

- **Slake** is a multilingual medical VQA dataset constructed from radiology images, including X-ray, CT, and MRI, with annotations provided in both English and Chinese, and emphasizes visual grounding with well-aligned question–answer pairs [19].
- **PathVQA** focuses on pathology images, containing high-resolution histopathological slides with questions that require fine-grained visual reasoning and domain-specific knowledge [7].
- **VQA-RAD** is a radiology VQA benchmark built upon clinical radiology images, where questions often involve abnormality detection, modality recognition, and anatomical reasoning; it is relatively small but widely used for evaluating medical reasoning capability [12].
- **PMC-VQA** is a large-scale dataset derived from biomedical literature, consisting of hundreds of thousands of image–question–answer pairs spanning multiple modalities (e.g., radiology, microscopy, charts) [34].

We employ a part of their training set to construct our ***MedVR-3K*** and their test set for evaluation, and the detailed statistics of each dataset are summarized in Table 5. In particular, after quality control, we adopted 746, 721, 707, 1303 samples from Slake, PathVQA, VQA-RAD, and PMC-VQA respectively.

Evaluation Metrics. We adopt two complementary evaluation metrics to assess model performance under different question formats. (i) Accuracy: For multiple-choice questions, we use exact-match

accuracy, which measures whether the predicted answer $\hat{y}$ matches the ground-truth label $y^\star$:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{y}_i = y_i^\star), \tag{20}$$

where $\mathbb{I}(\cdot)$ is the indicator function. This metric is suitable for structured questions with a predefined set of candidate answers. (ii) Correctness: For free-form questions, exact matching is often overly restrictive due to linguistic variability. Therefore, we adopt a semantic correctness metric based on an independent and proprietary MLLM evaluator. Specifically, given a generated answer $\hat{y}$ and ground-truth answer $y^\star$. In practice, we employ the GPT 5.4 owing to its strong text information processing ability. Formally, the evaluator model $\mathcal{E}$ produces a score:

$$\text{Correctness} = \mathcal{E}(\hat{y}, y^\star) \in [0, 1], \tag{21}$$

which reflects the semantic alignment between the prediction and the reference.

B Derivation of the Reward-Induced Decision Threshold

We provide a derivation of the reward-induced decision threshold discussed in Sec. 3.2. At the router level, the model chooses between answering directly and declaring `UNCERTAIN`, which triggers multi-agent collaboration. We consider an idealized decision setting in which $p(x, q)$ is calibrated and the router selects the action with the highest expected certainty-aware reward.

Under the certainty-aware reward in Eq. (5), the expected utilities of the two actions are

$$U_{\text{ans}} = p(x,q) - (1 - p(x,q))\beta, \qquad U_{\text{def}} = p(x,q)\alpha. \tag{22}$$

The Bayes-optimal action under this reward selects the action with higher expected utility:

$$U_{\text{ans}} \geq U_{\text{def}}. \tag{23}$$

Substituting the above utilities, we obtain

$$p(x,q) - (1 - p(x,q))\beta \geq p(x,q)\alpha. \tag{24}$$

Rearranging terms gives

$$p(x,q)(1 + \beta - \alpha) \geq \beta, \tag{25}$$

which yields the reward-induced threshold

$$p(x,q) \geq \frac{\beta}{1 - \alpha + \beta}. \tag{26}$$

Therefore, the corresponding Bayes-optimal decision rule under the idealized calibrated setting is

$$a^*_{\text{Bayes}}(x,q) = \begin{cases} \text{ANS}, & p(x,q) \geq \tau_{\text{Bayes}}, \\ \text{DEF}, & p(x,q) < \tau_{\text{Bayes}}, \end{cases} \qquad \tau_{\text{Bayes}} = \frac{\beta}{1 - \alpha + \beta}. \tag{27}$$

This threshold should be interpreted as a theoretical utility indifference point induced by the reward parameters α and β, rather than as the exact decision boundary of the GRPO-trained router. In the actual model, GRPO learns a policy $\pi_{\theta^\star}(a \mid x, q)$, whose routing boundary is implicitly determined by

$$\pi_{\theta^\star}(\text{ANS} \mid x, q) = \pi_{\theta^\star}(\text{DEF} \mid x, q). \tag{28}$$

Consequently, the empirical routing boundary may differ from τ_{Bayes} due to imperfect calibration, finite training data, optimization dynamics, and the parameterization of the learned policy.

C Additional Analysis

Routing Validation. To verify that the learned router agent can exploit meaningful sample-level uncertainty rather than merely changing the average collaboration rate, we compare ***SMART-Med*** * against two extreme policies (always-single, always-collaborate) using the same fine-tuned backbone and a random-routing control that replaces the learned router with an independent Bernoulli decision at a matched collaboration rate. Table 6(a) shows ***SMART-Med*** * achieves the best accuracy on both MC and FF while issuing only 1.2 calls on average, outperforming both extremes; Table 6(b) shows random routing achieves far lower accuracy than learned routing despite a comparable single-agent rate, indicating the router identifies *which* queries benefit from collaboration rather than just how often to invoke it.

Table 6: Routing validation with a matched fine-tuned backbone. (a) Learned routing vs. two extreme policies; (b) learned routing vs. random routing at a matched collaboration rate.

Model	MC	FF	Avg. Calls
(a) Routing strategy comparison			
Single agent*	76.2	65.4	1.0
Dense collaboration*	77.5	66.8	4.0
***SMART-Med** * (Ours)	**78.1**	**68.5**	**1.2**
(b) Random vs. learned routing (% single-agent)			
Random routing ($p = 0.5$)	65.0 (49.5%)	62.2 (48.5%)	2.0
Uncertainty routing (Ours)	**78.1** (85.5%)	**68.5** (78.5%)	**1.2**

Risk–Coverage Analysis. We further quantify calibration quality via the Area Under the Risk–Coverage Curve (AURC; lower is better). Table 7 shows certainty-aware GRPO reduces overall AURC from 0.2669 to 0.2334 while improving full-coverage accuracy from 66.0% to 71.5%, confirming that our designed GRPO fine-tuning improves calibration and accuracy jointly.

Table 7: Risk–coverage analysis before/after certainty-aware GRPO.

Model	AURC ↓	Acc.@100% ↑
SMART-Med	0.2669	66.0%
***SMART-Med** *	**0.2334**	**71.5%**

False-Certainty Safety Analysis. The false-certainty is a typical failure mode in MedVQA. Table 8 cross-tabulates the router's certainty token against correctness before and after GRPO: the false-certainty rate (Incorrect+CERTAIN) drops from 28.8% to 8.6%, while incorrect predictions correctly flagged `UNCERTAIN` rise from 5.3% to 15.7%, showing that certainty-aware GRPO specifically suppresses overconfident errors and defers the remaining difficult cases to collaboration, which can then correct them (concrete examples can be found in the Appendix D).

Table 8: Certainty–correctness confusion matrices before/after certainty-aware GRPO.

Correctness	Before GRPO		After GRPO	
	CERTAIN	UNCERTAIN	CERTAIN	UNCERTAIN
Correct	63.4%	2.5%	71.5%	4.1%
Incorrect	28.8%	5.3%	8.6%	15.7%

Table 9: Additional robustness analyses. (a) Paired significance tests of ***SMART-Med** * vs. baselines (five runs). (b) Independent judge (Gemini 3 Pro) vs. GPT-5.4 for reasoning-quality scores.

(a) Significance vs. baselines

Comparison	p-value
vs. Reconcile	0.0025
vs. ColaCare	0.0043
vs. MedAgent	0.0137
vs. MDAgent	0.0341
vs. MOA	0.0860

(b) Judge comparison

Method	GPT-5.4	Gemini 3 Pro
MDAgent	2.75	3.12
MedAgent	3.20	2.94
SMOA	3.38	3.35
MOA	3.31	3.42
ColaCare	3.22	3.57
SMART-Med	3.11	3.20
***SMART-Med** *	**3.50**	**3.73**

Statistical Significance. We perform paired significance tests using per-question performance of ***SMART-Med** * against each baseline across the five runs, reported in Table 9(a). The improvements over Reconcile ($p = 0.0025$) and ColaCare ($p = 0.0043$) are significant at the 1% level, and the improvements over MedAgent ($p = 0.0137$) and MDAgent ($p = 0.0341$) are significant at the 5% level. The comparison with MOA does not reach significance ($p = 0.0860$), which is consistent with the relatively small performance gap between the two methods observed in Table 1. Overall, the pairwise comparisons are statistically significant at different salient levels, supporting that the accuracy gains of ***SMART-Med** * are statistically salient against most of baseline MAS models.

Extra Judge Evaluation. To verify the reasoning quality improvements of our reasoning quality reward across different judge models, we additionally score the same 100 sampled responses with Gemini 3 Pro, a model that is completely independent of both ***MedVR-3K*** construction and reward-model training, and report the overall average score (across the Logic Consistency, Factual Correctness, and Image Alignment dimensions used in Table 3) under both judges in Table 9(b). The ranking is largely preserved across the two judges, and ***SMART-Med*** * remains the top-scoring method under both GPT-5.4 (3.50) and Gemini 3 Pro (3.73). Notably, Gemini 3 Pro assigns ***SMART-Med*** * an even higher score than GPT-5.4 does, and the relative ordering of the multi-agent baselines (e.g., ColaCare and MOA scoring higher under Gemini 3 Pro than under GPT-5.4) shifts somewhat but does not change the overall conclusion. This cross-judge consistency indicates that the reasoning-quality advantage of ***SMART-Med*** * reflects a genuine property of the generated reasoning traces rather than a preference specific to GPT-5.4.

D Examples of our *SMART-Med*

While the quantitative results in Sec. 3.2 and the efficiency analysis in Fig. 4 demonstrate the efficiency gains of ***SMART-Med*** on straightforward cases, it is equally important to examine cases where the router's initial assessment is uncertain or incorrect, and where multi-agent collaboration successfully corrects the prediction. In this section, we present detailed inference traces contrasting ***SMART-Med*** against baseline multi-agent systems (MDAgent, MOA, and MedAgent where available) on two representative difficult cases from our evaluation. These examples illustrate not only that uncertainty-aware routing avoids unnecessary collaboration on easy cases, but also that collaboration is selectively invoked and effective precisely when the initial prediction is ambiguous or erroneous.

Example 1: Pathology Question on Inflammatory Infiltrate Characterization.

Question: Is the immediate reaction characterized by an inflammatory infiltrate rich in eosinophils, neutrophils, and T cells?

Ground Truth: NO

SMART-Med: • Initial assessment → certainty=UNCERTAIN → answer=YES (incorrect)
• Router recruits expert: Pathology
• Specialist analysis → Moderator aggregation
• Final decision=NO **(correct)**
• Calls=4; tokens_in/out=1468/355; time=34.75s

MDAgent: • Difficulty=intermediate → recruit=[Pulmonologist, Allergist/Immunologist, Pathologist]
• Specialist analyses and repeated inter-agent consultation
• Final decision=YES **(incorrect)**
• Calls=34; tokens_in/out=8184/4464; time=303.91s

MOA: • Three homogeneous agent analyses → moderator aggregation
• Final decision=NO **(correct)**
• Calls=4; tokens_in/out=2093/1728; time=123.73s

MedAgent: • Domain classification → domains=[Rheumatology, Immunology, Allergy]
• Domain analyses → synthesis/consensus
• Final decision=NO **(correct)**
• Calls=12; tokens_in/out=6067/2738; time=183.68s

Example 2: Multiple-Choice Question on Muographic Map Interpretation.

Question: What do the empty white stars indicate in the map of the measurement site?
A: Selected cross-sections.
B: Locations of lahar observation stations.
C: Selected angular regions.
D: Locations of muographic observation stations.

Ground Truth: B

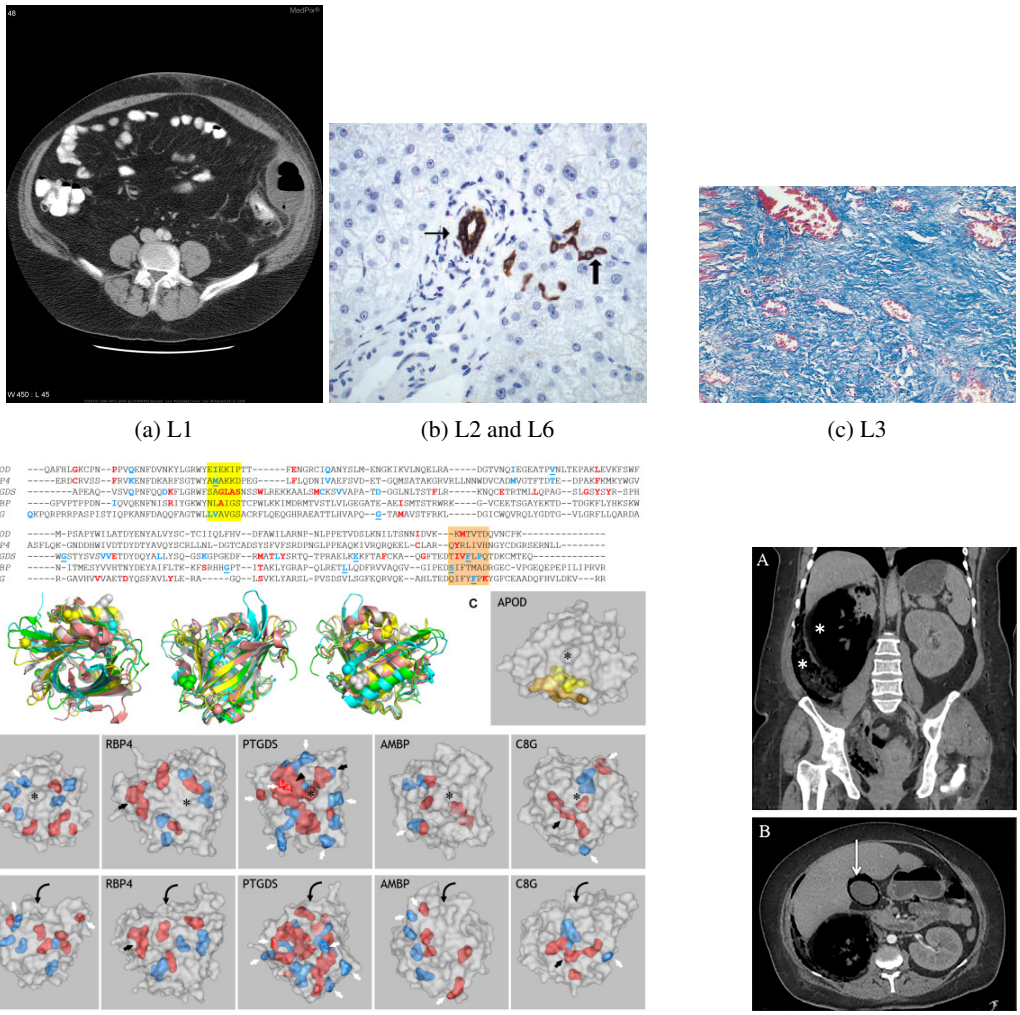
SMART-Med: • Initial assessment → certainty=UNCERTAIN → answer=B (correct)
• Router recruits expert: General Practice
• Specialist analysis → Moderator aggregation
• Final decision=B **(correct)**
• Calls=4; tokens_in/out=2231/394; time=38.34s

MDAgent: • Difficulty=intermediate → recruit=[Urologist, Pathologist, Oncologist]
• Specialist analyses and inter-agent consultation
• Final decision=B **(correct)**
• Calls=13; tokens_in/out=5705/1851; time=113.65s

MOA: • Three homogeneous agent analyses → moderator aggregation
• Final decision=D **(incorrect)**
• Calls=4; tokens_in/out=2657/1446; time=81.97s

MedAgent: • Domain classification → domains=[Field0, Field1, Field2]
• Domain analyses → synthesis/consensus
• Final decision=B **(correct)**
• Calls=13; tokens_in/out=8930/2430; time=164.17s

Discussion. These two examples demonstrate that ***SMART-Med***'s router recognizes its own uncertainty rather than returning an overconfident wrong answer, and invokes collaboration only in these flagged cases. In Example 1, the router's initial prediction (YES) is incorrect, but the system correctly flags its uncertainty and recruits specialist reasoning, which successfully corrects the error to the ground-truth answer (NO). Notably, MDAgent deploys dense collaboration (34 calls) yet still produces an incorrect final answer, while MOA and MedAgent achieve correctness at substantially higher computational cost (4–12 calls vs. ***SMART-Med***'s 4 calls, but with 1.4–6.1$\times$ more tokens). In Example 2, ***SMART-Med***'s initial uncertain prediction happens to be correct and is confirmed by collaboration with minimal overhead (4 calls, 38.34s), whereas MOA—despite also using 4 calls—produces an incorrect answer (D), and MDAgent and MedAgent require 3–4$\times$ more calls and time to reach the correct conclusion. These cases illustrate that dense collaboration alone does not guarantee correctness; in contrast, uncertainty-aware routing targets collaboration precisely where it is most needed, improving both safety and efficiency by reducing overconfident errors while maintaining or improving diagnostic accuracy.



(a) L1 (b) L2 and L6 (c) L3

(d) L4 (e) L5

Figure 5: The medical images used in the Listing 1 to Listing 6. Notably, the image (b) is used for two queries, a multi-choice question (L2) and a free-form question (L6).

E Showcases of *MedVR-3K*

We present and analyze samples from ***MedVR-3K*** to showcase the diversity, reasoning quality, and the reliability of multi-dimensional annotations. The corresponding images are shown in Fig. 5.

The selected cases span multiple modalities and knowledge domains, including histopathology (Listings 2, 3, and 6), radiology (Listings 1 and 5), and structural biology (Listing 4), demonstrating the broad coverage of ***MedVR-3K***. Importantly, the dataset captures a wide spectrum of reasoning behaviors. For instance, Listing 2 shows a case where the model reaches the correct conclusion (CK7 staining of biliary structures) but introduces a domain-specific factual error, reflecting realistic partial understanding. Listing 1 illustrates a low-quality response with minimal reasoning and weak image grounding despite a correct final answer, while Listing 3 provides a strong positive example with accurate domain knowledge and precise visual alignment in identifying collagen under trichrome staining. Listing 4 further highlights cases where answers are correct but supported by weak or generic reasoning, whereas Listing 5 demonstrates high-quality spatial reasoning and robust image grounding in a radiological context. Finally, Listing 6 captures a subtle semantic discrepancy, where the model's description (bile duct and ductules) is close but not strictly equivalent to the ground-truth concept (bile duct cells and canals of Hering), reflecting the fine-grained nature of medical terminology.

Meanwhile, for Listing 5, we also compare it with the doctor evaluation (i.e., Listing 7). Both responses reach the exact same conclusion and share the same key reasoning: the gas is located outside the necrotic right kidney within the retroperitoneal space, while the gallbladder, left kidney, and stomach are all anatomically incorrect locations. The main differences lie in style and structure: the physician's response is more detailed and pedagogical, presenting a clear step-by-step elimination of each distractor without needing to cross-reference Fig. 5(e)-B. Although the MLLM's response is more direct in reaching the answer, it still contains the key analytical reasoning chain (e.g., step-by-step elimination), demonstrating that our constructed dataset is well-equipped with human-doctor-level reasoning chains.

For the associated quality annotations in ***MedVR-3K***, including logical consistency, factual correctness, and image alignment—closely match expert judgment across these varied scenarios. They successfully distinguish between correct answers with flawed reasoning (Listings 2 and 4), unsupported predictions (Listing 1), high-quality reasoning (Listings 3 and 5), and semantically imprecise outputs (Listing 6). This demonstrates that ***MedVR-3K*** not only provides diverse and clinically relevant VQA samples, but also enables nuanced and reliable evaluation of model behavior. Overall, ***MedVR-3K*** serves as a challenging and well-calibrated benchmark for advancing medical multimodal reasoning.

Listing 1: Example 1


```
{
  "question": "Are air fluid levels depicted in this image?",
  "image-path": "synpic28718.jpg",
  "response": "I'm not 100% sure for my analysis to this medical image and the
  given question. But, I can guess the answer is YES. \n\n<answer>YES</answer>",
  "label": "YES",
  "logic_consistency": 1,
  "factual_correctness": 3,
  "image_alignment": 1,
  "difficulty": 2,
  "judgment": "The assistant provides essentially no reasoning and explicitly
  states uncertainty, then gives a bare YES answer, so the conclusion is
  unsupported and logically weak. Factual correctness is difficult to verify from
   the response alone; the image does appear to show intraluminal gas-fluid
  interfaces, so the final answer may be correct, but the lack of explanation and
   hedging prevent a higher score. Image alignment is poor because the assistant
  does not reference any specific visual features or regions in the CT image. The
   task itself is a basic image-identification question requiring limited medical
   knowledge, so difficulty is low."
}
```

Listing 2: Example 2


```
{
    "question": "Are bile duct cells and canals of Hering stained here with an
    immunohistochemical stain for cytokeratin 7?",
    "image-path": "train_0422.jpg",
    "response": "The image displays a histological section with cells stained to
    reveal specific structures. The presence of darkly stained, duct-like
    structures (indicated by arrows) amidst a background of smaller, round, blue-
    stained nuclei is consistent with the appearance of bile ducts and canals of
    Hering under immunohistochemical staining. Cytokeratin 7 (CK7) is a well-known
    marker that stains bile duct epithelium and canals of Hering, which are
    typically positive for CK7. The staining pattern in the image, with strong,
    specific labeling of the ductal structures, aligns with the expected
    immunohistochemical profile for CK7. The surrounding cells appear to be
    hepatocytes, which are also known to express CK7, but the most prominent
    staining is in the ductal structures. Given this evidence, the staining is
    consistent with an immunohistochemical stain for cytokeratin 7.<answer>YES</
    answer>"
    "label": "YES",
    "logic_consistency": 3,
    "factual_correctness": 2,
    "image_alignment": 4,
    "difficulty": 2,
    "judgment": "The response gives a coherent rationale linking brown
    immunostaining of duct-like structures to CK7-positive bile ducts/canals of
    Hering, so the logic is generally understandable but somewhat overstated and
    not fully rigorous. Image grounding is reasonably good because it refers to the
     dark brown ductal structures indicated by arrows against blue nuclei, which
    matches the visual features. However, factual accuracy is reduced by the claim
    that surrounding hepatocytes are also known to express CK7; in routine liver
    histology, hepatocytes are generally CK7-negative, while bile duct epithelium
    and canals of Hering are CK7-positive. That is a significant pathology-specific
     error, though the final YES answer is likely correct. The question itself is a
     basic histopathology/IHC identification task requiring limited specialist
    knowledge, so difficulty is low-to-moderate."
}
```

Listing 3: Example 3


```
{
  "question": "Is Collagen stained blue by the trichrome stain?",
  "image-path": "train_0473.png",
  "response": "In histological staining, the trichrome stain is commonly used to
  differentiate collagen from other tissue components. In this stain, collagen
  fibers are typically stained blue or green, while muscle fibers appear red or
  pink, and nuclei are often counterstained with a different color (e.g., purple
  or black).\n\nIn the provided image, the background matrix appears
  predominantly blue, and there are scattered pinkish-purple structures that
  resemble nuclei or other cellular components. The blue-stained areas have a
  fibrous, wavy appearance consistent with collagen fibers.\n\nTherefore, based
  on the visual evidence and standard histological knowledge, collagen is stained
   blue by the trichrome stain.\n\n<answer>YES</answer>",
  "label": "YES",
  "logic_consistency": 5,
  "factual_correctness": 5,
  "image_alignment": 5,
  "difficulty": 3,
  "judgment": "The assistant provides a clear, step-by-step explanation linking
  trichrome staining principles to the image features, correctly identifying blue-
  stained fibrous areas as collagen. The description of other components (pink
  nuclei, background) aligns with standard histology. The response is factually
  accurate and demonstrates precise visual analysis of the image, though the
  question itself is moderate in complexity, requiring basic histological
  knowledge and single-step reasoning."
}
```

Listing 4: Example 4


```
{
  "question": "What do the colored spheres in Figure B represent? A: Ligand
  pockets. B: Beta-barrels. C: Type-I residues. D: SDP residues.",
  "image-path": "PMC8420045_F3.jpg",
  "response": "I can discern various colored spheres across the protein structures
   shown in Figure B. These spheres seem to highlight specific types of residues
  based on their chemical characteristics.\n\nFrom the image:\n- The spheres are
  shown in different colors (such as red, blue, and white), which could indicate
  specific features of the residues.\n- The context of the surrounding figures (
  which display proteins and their structural features) suggests that the colored
   spheres are likely highlighting specific residues that are crucial for
  particular interactions or functions.\n\nGiven the evidence in the image and
  the provided options, the most plausible conclusion is that these spheres
  represent **Type-I residues**, as they are typically associated with specific
  amino acid positions critical for structural functions in protein analysis.\n\n<
  think>Based on the visualization and the context of protein analysis, the
  colored spheres in Figure B most likely represent Type-I residues.</think><
  answer>C</answer>",
  "label": "C",
  "logic_consistency": 5,
  "factual_correctness": 5,
  "image_alignment": 4,
  "difficulty": 3,
  "judgment": "The response correctly identifies the answer as C (Type-I residues)
  , matching the label, so factual correctness is full. The internal reasoning is
   coherent and non-contradictory, though it lacks strong evidence from the
  figure; logic consistency is still high. Without seeing the image, alignment is
   judged as reasonable (4) because the description of colored spheres is
  plausible for a protein figure, and the correct choice suggests good visual
  matching. Difficulty is moderate as the question requires knowledge of protein
  structural features but is not extremely specialized."
}
```

Listing 5: Example 5


```
{
  "question": "What is the bottom asterisk pointing to in the image A? A:
  Intramural gas in the gallbladder B: Gas extending through the retroperitoneum
  C: Extensive parenchymal destruction of the left kidney D: Inflammation in the
  stomach.",
  "image-path": "PMC8442220_f2.jpg",
  "response": "<think>The coronal CT image A shows the patient's right side on the
   left of the image. There is a markedly abnormal right renal area with
  extensive gas replacing much of the renal parenchyma. The upper asterisk is
  within this gas-filled destructive renal region. The bottom asterisk points to
  streaky/mottled gas tracking outside the kidney along the right flank/
  retroperitoneal space. This is not the gallbladder (seen in image B with arrow),
   not the left kidney, and not the stomach.</think><answer>B</answer>",
  "label": "B",
  "logic_consistency": 4,
  "factual_correctness": 4,
  "image_alignment": 5,
  "difficulty": 3,
  "judgment": "The assistant gives a clear rationale: it distinguishes the lower
  asterisk from the upper renal gas collection, notes that the lower marker
  indicates gas tracking outside the kidney along the flank/retroperitoneum, and
  explicitly rules out the gallbladder, left kidney, and stomach. This is
  logically coherent and matches the visual appearance of image A. The answer
  choice B is medically appropriate; the image shows emphysematous change with
  gas extending into the retroperitoneal/perinephric region rather than
  gallbladder gas, left renal destruction, or gastric inflammation. Image
  grounding is strong because the response localizes the finding to the specific
  lower asterisk and relates it to surrounding anatomy. A perfect logic/factual
  score is not warranted because the reasoning is concise rather than exhaustive,
   and it references image B to exclude the gallbladder even though the question
  asks only about image A. The question itself is of moderate difficulty,
  requiring standard radiologic anatomy and pathology recognition rather than
  highly advanced inference."
}
```

Listing 6: Example 6


```
{
  "question": "What are stained here with an immunohistochemical stain for
  cytokeratin 7?",
  "image-path": "train_0422.jpg",
  "response": "<think>The image is a micrograph of liver tissue where an
  immunohistochemical stain (brown) has been applied, likely for cytokeratin 7.
  The thin arrow points to a distinct, lumen-containing structure within a portal
   tract, which is morphologically characteristic of a bile duct. The thick arrow
   points to smaller, irregular clusters of stained cells at the periphery of the
   portal tract, representing bile ductules. Both of these biliary epithelial
  structures are known to be positive for cytokeratin 7, while the surrounding
  hepatocytes are negative.</think><answer>A bile duct (thin arrow) and bile
  ductules (thick arrow)</answer>",
  "label": "BILE DUCT CELLS AND CANALS OF HERING",
  "logic_consistency": 5,
  "factual_correctness": 3,
  "image_alignment": 3,
  "difficulty": 3,
  "judgment": "The response is internally logical and clearly describes the two
  arrow labeled structures. However, the label specifies 'bile duct cells and
  canals of Heringm' while the answer says 'bile duct and bile ductules', bile
  ductules are not exactly the same as canals of Hering (the latter are
  transitional channels between canaliculi and ductules), so factual correctness
  is slightly reduced. Without seeing the actual image, the alignment score is a
  moderate 3, assuming the arrows might not perfectly match the labeled
  structures. Difficulty is average (3) because CK7 staining in liver is a
  standard histopathology concept."
}
```

Listing 7: Human Doctor Evluation to Example 5

```
{
  "question": "What is the bottom asterisk pointing to in the image A? A:
  Intramural gas in the gallbladder B: Gas extending through the retroperitoneum
  C: Extensive parenchymal destruction of the left kidney D: Inflammation in the
  stomach.",
  "image-path": "PMC8442220_f2.jpg",
  "doctor": "(1) Image Localization: Figure A is a coronal CT image, with the left
   side of the image corresponding to the patient's right side. A large area of
  air-containing low-attenuation shadow can be seen in the right renal region,
  indicating renal parenchymal necrosis with gas accumulation. The upper asterisk
   marks the intrarenal gas and necrosis, while the lower asterisk marks the
  extrarenal gas shadow located in the retroperitoneal space.

  (2) Systematic elimination of distractors!
  A (Intramural gas in the gallbladder wall): The gallbladder is located in the
  right upper quadrant liver bed area, which does not match the anatomical
  location of the lower asterisk. The arrow in Figure B merely serves as a
  reference for the gallbladder and can be excluded without cross-referencing
  between images.
  C (Extensive parenchymal destruction of the left kidney): The left kidney is
  located on the right side of the image, with essentially normal morphology and
  density, showing no gas accumulation or extensive destruction. The lesion is
  entirely on the right side and can therefore be excluded.
  D (Gastric inflammation): The main body of the stomach is in the left upper
  quadrant, posterior to the liver; the gastric antrum is near the right upper
  quadrant, adjacent to the pancreatic head. This does not correspond to the
  right-sided retroperitoneal location of the lower asterisk and can be excluded.

  (3) Conclusion: The gas indicated by the lower asterisk extends outward from the
   necrotic right kidney into the retroperitoneal space, corresponding to option
  B.",
}
```

F Prompts Design

Prompts for Building *MedVR-3K*. ***MedVR-3K*** is a reasoning-centric dataset, and the reasoning quality is critical. To ensure the quality, we employ an independent judge MLLM (i.e., GPT 5.4) to evaluate the quality of each reasoning chain. We design the prompts for the judge model as below:

Prompts for the judge MLLM

```
You are an expert evaluator for multimodal medical QA tasks. Your role is to critically assess the
assistant's response to the provided User Input.

### User's Question
{question}

### Assistant Response
{assistant_response}

---

## Evaluation Rubrics

Evaluate the assistant's response on the following four dimensions. Each dimension is scored on a
1-5 scale.

### 1. Logic Consistency (1-5)
Assesses the coherence, structure, and validity of the reasoning chain.

| Score | Criteria |
|-------|----------|
| 1 | Severe logical fallacies, self-contradiction, circular reasoning, or complete lack of
reasoning |
```

```
| 2 | Flawed reasoning with significant gaps; conclusion does not logically follow from premises |
| 3 | Generally sound reasoning but lacks depth, contains minor leaps, or oversimplifies
complexity |
| 4 | Clear, structured reasoning that follows medical diagnostic/analytical standards |
| 5 | Flawless, step-by-step deductive reasoning; robust, complete, and clinically sound |

### 2. Factual Correctness (1-5)
Assesses the medical accuracy of the response.

| Score | Criteria |
|-------|----------|
| 1 | Critical hallucinations; dangerous or completely incorrect medical advice/claims |
| 2 | Significant factual errors regarding symptoms, anatomy, pathology, or treatment |
| 3 | Mostly correct but contains minor inaccuracies, outdated information, or imprecise
terminology |
| 4 | Factually correct throughout; no errors detected |
| 5 | Factually correct and precise; uses professional terminology accurately and appropriately |

### 3. Image Alignment (1-5)
Assesses how well the response utilizes and accurately interprets visual information.

| Score | Criteria |
|-------|----------|
| 1 | Completely ignores the image; describes features not present or contradicts visual evidence
|
| 2 | Mentions the image but misinterprets key visual features (e.g., wrong location, wrong
modality, incorrect identification) |
| 3 | Generic response that could apply to any similar image; lacks specific visual grounding |
| 4 | Correctly identifies and describes relevant visual features; demonstrates basic visual
grounding |
| 5 | Deep visual grounding; precisely correlates specific image regions with clinical reasoning;
demonstrates expert visual analysis |

### 4. Difficulty (1-5)
Assesses the complexity of the question itself (not the answer quality).

| Score | Criteria |
|-------|----------|
| 1 | Simple fact retrieval or common knowledge; no medical expertise required |
| 2 | Basic identification task; requires minimal medical knowledge |
| 3 | Moderate complexity; requires standard medical knowledge and single-step reasoning |
| 4 | Complex case; requires multi-step reasoning, integration of multiple concepts, or specialist
knowledge |
| 5 | Highly complex; requires integrating multiple modalities, rare/specialized knowledge, or
advanced differential diagnosis |

---

## Strict Evaluation Guidelines

1. **Be conservative**: A score of 5 is reserved for "perfect" or "exemplary" responses. Use 4 for
"good but not perfect."

2. **Penalize hallucinations heavily**: Any fabricated information should result in a Factual
Correctness score smaller than 2.

3. **Guess vs. reasoning**: If the reasoning is missing, flawed, or insufficient, but the final
answer is correct (guessed), cap the Logic Consistency score at 2.

4. **Image requirement**: A response that fails to meaningfully engage with the image (when visual
input is essential) cannot receive an Image Alignment score above 3.

5. **Response completeness**: If the response is incomplete or refuses to answer without
justification, scores should reflect this deficiency.

6. **Evidence-based judgment**: All scores must be justified with specific evidence from the
assistant's response.

---

## Output Format

Output ONLY a valid JSON object. Do not include markdown formatting, code blocks, or explanations
outside the JSON.

{{
    "scores": {{
        "logic_consistency": <int 1-5>,
        "factual_correctness": <int 1-5>,
        "image_alignment": <int 1-5>,
```

```
        "difficulty": <int 1-5>
    }},
    "critique": "<brief, evidence-based justification for your scores>"
}}
```

Prompts for Building *SMART-Med*. In our method, there are three kinds of roles, undertaking different tasks: (1) Router: the agent for initial assessment, certainty check and expert recruitment; (2) Specialist: the agent is assigned with specific tasks and collaboration patterns in the multi-agent team; (3) Moderator: the agent aims to summarize the analysis reports from multiple agents. Their system prompts, role prompts and task prompts are designed as follows:

Prompts for the router agent

```
# System Prompt
You are a medical resource allocation router. Your task is to allocate the medical question to the
most appropriate specialist agents efficiently and effectively.

## Task
Please recruit medical experts for this question: <image> {question}

## Instruction
1. Recruit maximally 5 specialist roles relevant to answer this question. Reference roles:
[Radiologist, Pathologist, Surgeon, Oncologist, Endocrinologist, Geneticist, Cardiologist,
Neurologist]
2. Generate a brief task description for each role explaining their job in the collaboration to
correctly answer this question.
3. For role descriptions, using **|** to seperate different roles.

## Output Format
<role>[Role1, Role2, ...]</role>\n<role_desc>[Role1: Description | Role2: Description |
...]</role_desc>

## Example
<role>Radiologist, Oncologist</role>\n<role_desc>Interprets CT images to detect lung masses |
Evaluates malignancy and treatment strategy</role_desc>
```

Prompts for the specialist agent

```
# System Prompt
Your task is {role_description}, and your job is {task_description}.

## Task
Analyze the provided medical image and answer the following {answer_type} question: <image>
{question}
## Instructions
1. **Visual Analysis**: Thoroughly examine the image and identify all anatomical structures,
abnormalities, or features relevant to the question.
2. **Evidence-Based Reasoning**: Base your conclusion strictly on the visual evidence. Cite
specific observations that support your analysis.
3. **Uncertainty Annotation**: If the evidence is genuinely ambiguous or critical features are
obscured, return UNCERTAIN in your response.
## Output Format
<reason>[Your visual analysis citing specific
evidence]</reason><certainty>[CERTAIN/UNCERTAIN]</certainty><answer>[Your definitive best-effort
answer: {answer_type}]</answer>
.
## Examples

**High Confidence Case:**
<reason>The CT slice clearly shows a hyperdense, well-circumscribed lesion in the right lung base
with characteristics typical of a
hamartoma.</reason><certainty>CERTAIN</certainty><answer>Benign</answer>

**Low Confidence Case:**
<reason>The image shows a possible mass in the mediastinum, but the slice level appears to be
above the cardiac silhouette and the structure boundaries are poorly defined due to motion
artifact.</reason><certainty>UNCERTAIN</certainty><answer>Malignant</answer>

## Critical Rules
1. The `<certainty>` tag MUST be either `CERTAIN` or `UNCERTAIN` (no other values allowed).
```

```
2. The `<answer>` tag MUST always contain a specific {answer_type} value, regardless of certainty
level.
```

Prompts for the moderator agent

```
# System Prompt
You are a medical assistant who excels at summarizing and synthesizing based on multiple experts
from various domain experts.

## Context
**Image**: <image>
**Question**: {question}
**Answer Type**: {answer_type} (Your final answer MUST strictly conform to this type)

**Initial Assessment**:
{initial_assessment}

**Expert Reports**:
{report}

## Analysis Framework

### 1. Individual Assessment
For the given initial assessment and expert reports, evaluate:
- **Certainty Level**: CERTAIN or UNCERTAIN (from <certainty> tag)
- **Key Evidence**: Specific visual features cited from the image
- **Reasoning Quality**: Logical soundness of the inference chain
- **Risk of Hallucination**: Whether claims exceed visible evidence

### 2. Cross-Expert Comparison
- **Consensus Detection**: Identify agreements and disagreements across experts and the initial
assessment
- **Evidence Strength**: Which conclusion has stronger visual support?
- **Certainty Weighting**: Prioritize CERTAIN answers unless their evidence is weaker
- **Conflict Analysis**: If experts disagree, determine if the disagreement stems from:
    - Different interpretation of the same evidence → weigh reasoning quality
    - Different evidence focus → check which evidence is more relevant to the question
    - Fundamental ambiguity in the image → acknowledge irreducible uncertainty

### 3. Adjudication Rules (Apply in Order)
1. **Strong Consensus**: If there are more than two experts certainly agree with strong evidence,
adopt their answer
2. **Quality Override**: If UNCERTAIN experts have stronger evidence than CERTAIN ones → favor
evidence over certainty label

### 4. Meta-Evaluation (Critical Step)
Even if experts agree with the initial assessment, ask yourself:
- Is there a **plausible alternative interpretation** they missed?
- Does their conclusion **overreach the visual evidence**?
- Would a **different expert specialty** (e.g., radiologist vs. pathologist) disagree?
Finally, your task is to determine the answer of this question according to your analysis to the
initial assessment and expert reports.

## Output Requirements
**CRITICAL**: Your answer MUST be a valid {answer_type} value.

## Output Format
<reason>[Step-by-step adjudication: summarize expert positions, compare evidence quality, explain
your reasoning for selecting the final answer]</reason><answer>[Your final answer]</answer>
```

G Limitation and Social Impact

Limitation. In ***SMART-Med***, we employ local small-scale MLLMs as the agent. Although it can preserve the privacy, save the computational cost and latency, the overall performance is still bounded by the capacity of the underlying backbone model, which may limit its ability to handle extremely complex or rare medical cases compared to very large proprietary models. Therefore, how to deal with extremely complex or rare medical cases can be a potential future direction.

Social Impact. This work has important implications for the development of privacy-preserving and deployable medical AI systems. By enabling fully local inference and reducing reliance on cloud-based models, ***SMART-Med*** helps mitigate data privacy and regulatory concerns in clinical

environments. Moreover, the integration of uncertainty-aware decision-making can improve the reliability and safety of AI-assisted diagnosis by reducing overconfident errors. However, potential risks remain: incorrect or biased reasoning generated by the model could still mislead practitioners if used without proper oversight, and the dataset construction process may inherit biases from source data and teacher models. Therefore, ***SMART-Med*** should be deployed as a decision-support tool rather than a replacement for medical professionals, and future work should focus on improving robustness, fairness, and clinical validation.