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

Jiacheng Hou[1], Xueliang Cui[23], Dan Lu[5] , Ruxin Wang[34] *

[1]Tsinghua Shenzhen International Graduate School [2]Southern University of Science and Technology [3]Shenzhen MSU-BIT University [4] Shenzhen Institutes of Advanced Technology, CAS [5] West China Hospital, Sichuan University

NeurIPS 2026 · Presenter: Jiacheng Hou

Background

Multi-agent systems have become an effective paradigm for medical VQA reasoning

1. **Multi-agent systems (MAS):** several multimodal LLMs take different roles and collaborate through a predefined workflow to solve complex tasks

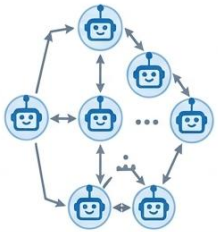
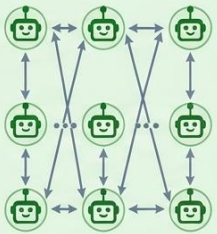
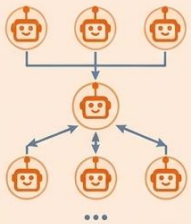
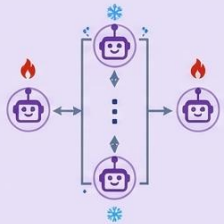
2. **Medical settings:** mimic the hospital workflow of triage → specialist consultation → final decision, as in MedAgents, MDAgents, and MMedAgent-RL

3. **Results:** multi-agent collaboration generally beats single models on medical VQA, for example raising PMC-VQA accuracy from 35.3% to 50.1%

Tang et al., 2024 (MedAgents); Kim et al., 2024 (MDAgents); Xia et al., 2026 (MMedAgent-RL); example data: paper Table 4

Motivation

Dense collaboration with cloud LLMs is costly and raises privacy risks

Method	(a) MDAgent	(b) MOA	(c) ColaCare	(d) MMedAgent-RL	(e) ***SMART-Med (Ours)***
Topology					
Paradigm	Multi-agent	Multi-agent	Multi-agent	Multi-agent	Adaptive
Base Model	Online	Online	Online	Online + Local Models	Completely Local Models
Training Objective	×	×	×	Answer-only	**Answer + Reasoning**

1. Dense collaboration: every question goes through multi-agent, multi-round interaction, which means high latency and many tokens for limited gains

2. Cloud LLMs: they rely on GPT, Gemini, and similar services, so patient images must be uploaded to external servers, which carries privacy and compliance risk

Source: paper Fig. 1; Zhu et al., 2025 (MedAgentBoard)

Research Question

Core question: can the model decide by itself when collaboration is needed?

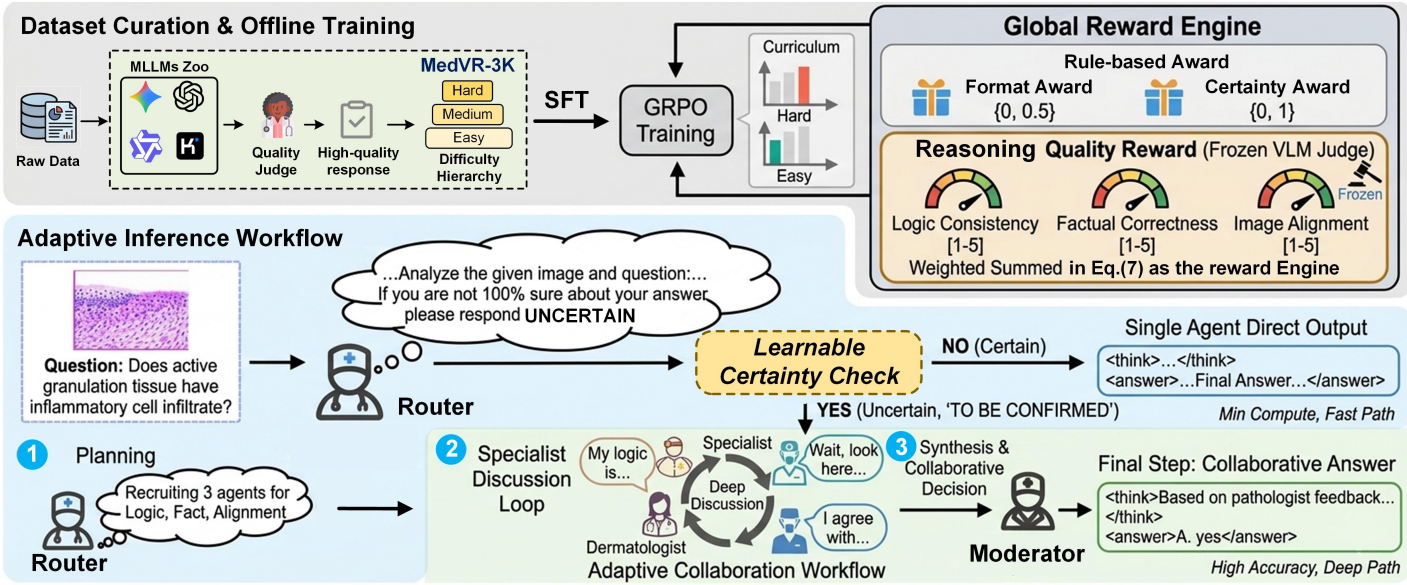
Key insight: Not every question needs a consultation. A single agent answers most questions reliably, and only uncertain ones justify convening experts. Collaboration should therefore be a learned decision, not a default behavior.

Main contributions

1. **SMART-Med:** the first fully local medical multi-agent system based on uncertainty-driven routing
2. **MedVR-3K:** 3,477 medical reasoning chains with difficulty levels and quality scores
3. **Dual-reward GRPO:** a confidence-aware accuracy reward plus a reward model for vision-language reasoning quality

Method

SMART-Med answers with a routing agent first and convenes experts only when uncertain



1. Routing: give an initial answer and flag whether it is certain → **2. Dispatch:** output certain answers directly; otherwise recruit up to 3 experts → **3. Decision:** the moderator summarizes the expert reports and gives the final answer

Source: paper Fig. 2

Method

GRPO jointly optimizes format, confidence calibration, and reasoning quality with three rewards

Format reward FMT: enforces the reason / certainty / answer structure

$$R_{\mathrm{FMT}}(o_i) = \mathbb{I}\left[\, o_i \text{ matches the output schema} \,\right]$$

Confidence-aware accuracy reward CAR: couples "is it confident" with "is it correct"

$$R_{\mathrm{CAR}}(o_i) = \mathbb{I}(y_i = y^\star)\left[c_i + (1 - c_i)\,\alpha\right] - \mathbb{I}(y_i \neq y^\star)\, c_i\, \beta$$

Reasoning quality reward RR: a frozen Qwen3-VL-2B scores the reasoning chain

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

CAR reward matrix

	Certain	Uncertain
Correct	+1	+α
Wrong	**−β**	0

If an answer may be wrong, the model should say "uncertain" and hand off to experts;
confidently wrong answers are penalized heavily

Total reward $\bar{R}_i = \lambda_{\mathrm{fmt}} R_{\mathrm{FMT}} + \lambda_{\mathrm{car}} R_{\mathrm{CAR}} + \lambda_{\mathrm{rr}} R_{\mathrm{RR}}$

Each input samples 6 responses; they are group-normalized to form the advantages

Shao et al., 2024 (GRPO); equations: paper Eq. (4)–(7)

Dataset

MedVR-3K: 3,477 stratified medical VQA items with reasoning chains

Collection
reasoning chains generated by 4 strong models, including GPT-5.4, for 4 benchmark training sets

→

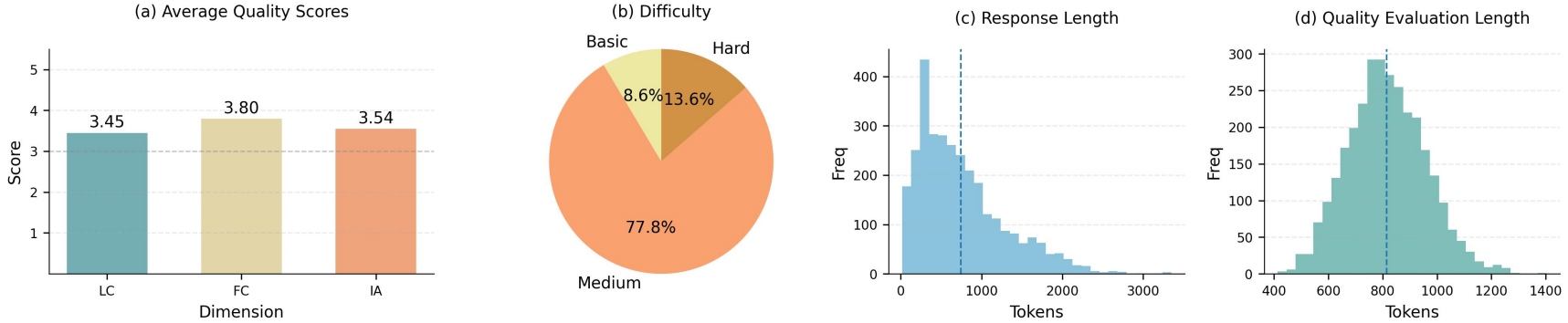
Filter 1
keep only reasoning chains that reach the correct final answer

→

Filter 2
score logic, factuality, and image alignment; apply a threshold per dimension

→

Difficulty labels
basic / medium / hard, supporting curriculum learning

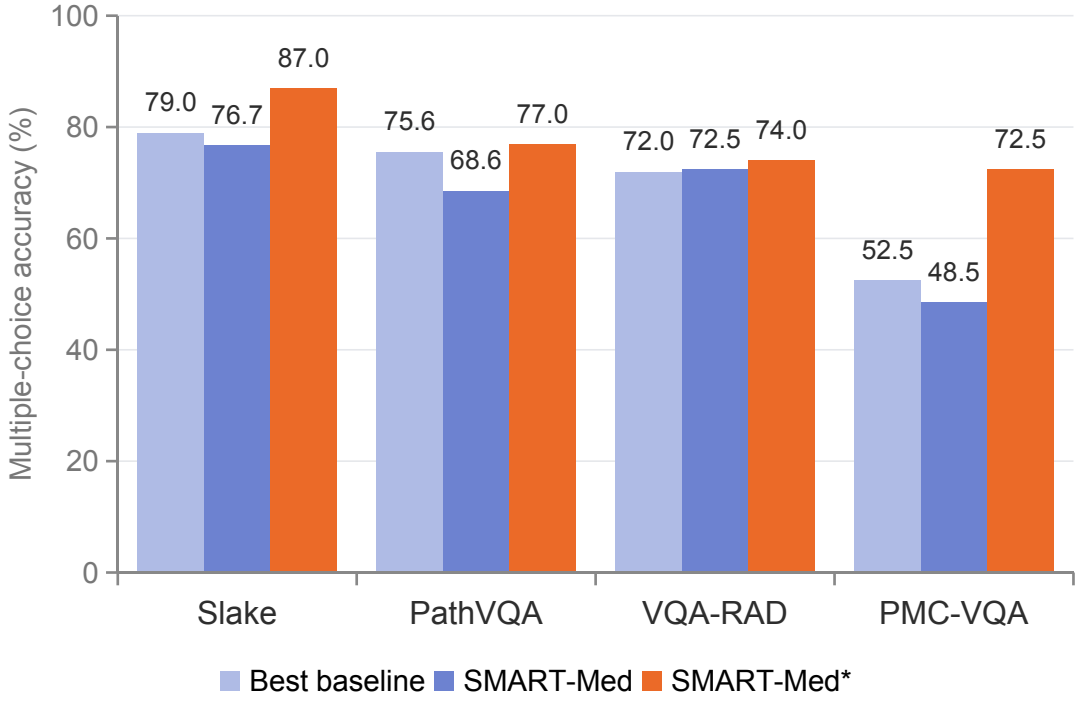


Average quality score ≈ 3.60 / 5 Sources: Slake 746 · PathVQA 721 · VQA-RAD 707 · PMC-VQA 1,303

Source: paper Fig. 3

Results

After fine-tuning, SMART-Med* ranks first on the multiple-choice questions in all 4 benchmarks



Zero-shot: untrained SMART-Med already matches dense MAS

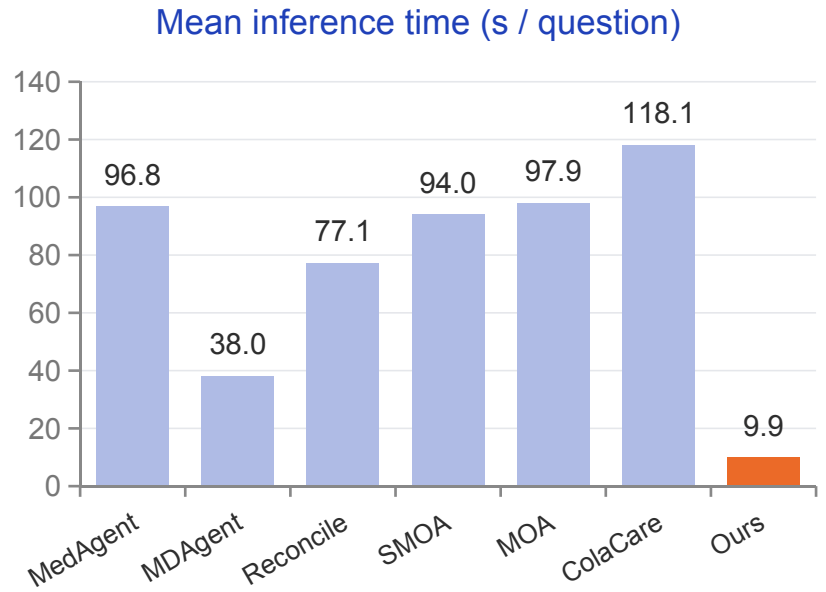
After fine-tuning: SMART-Med* exceeds the best baseline by +8.0 / +1.4 / +2.0 / +20.0 points

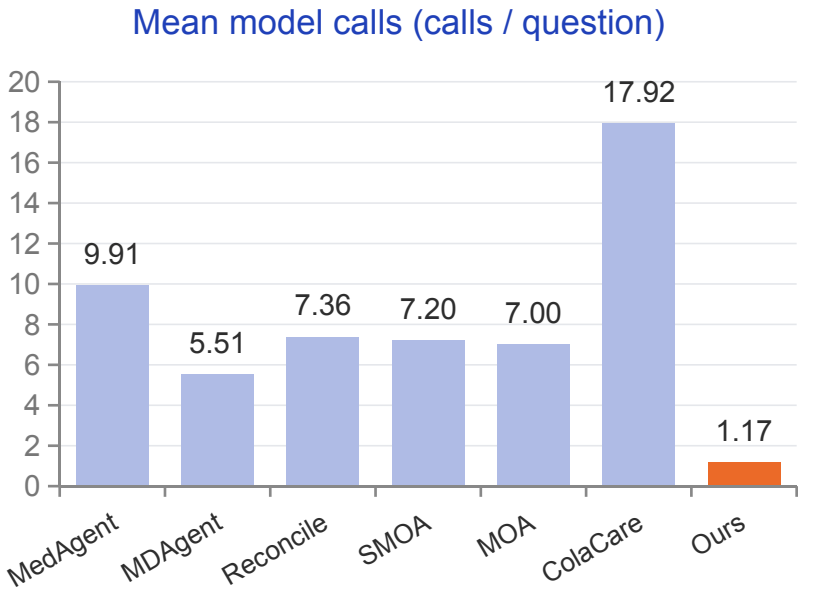
Open-ended questions: also the best, with Slake 84.3%, PathVQA 58.5%, and VQA-RAD 69.4%

Data: paper Table 1 (mean of 5 independent runs); all multi-agent methods use Qwen3-VL-4B-Instruct as the backbone

Results

Only 9.9 s and 1.17 calls per question, 3.8× faster than the fastest baseline





3.8× faster than MDAgent

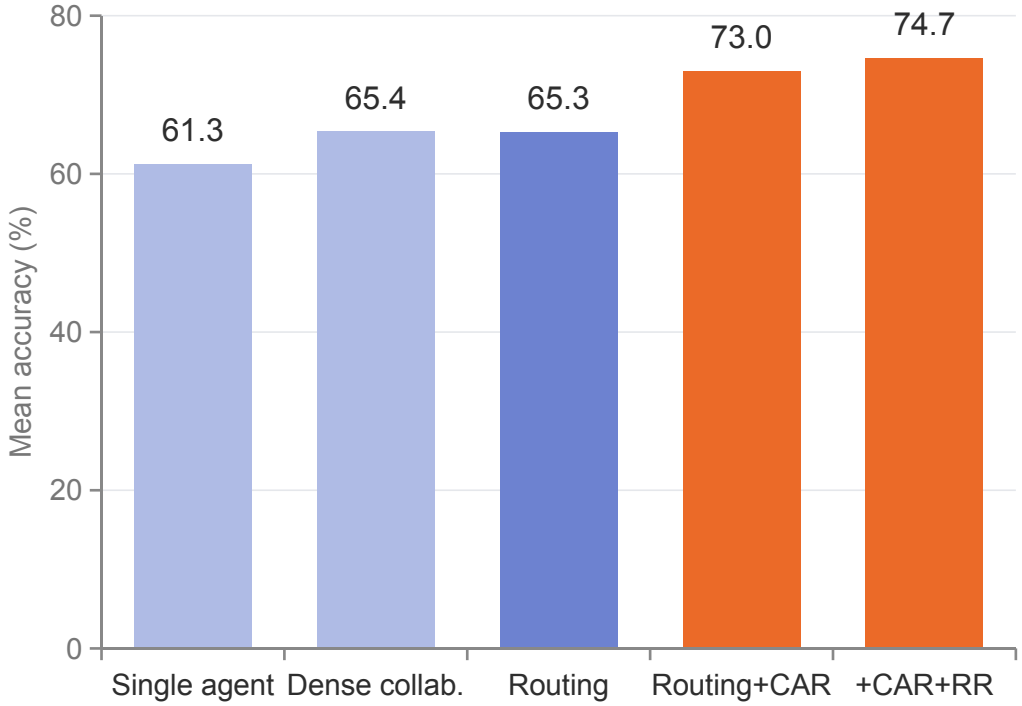
11.9× faster than ColaCare

−72% input tokens

Data: paper Fig. 4 (100 random samples); orange marks the proposed method

Results

Routing matches dense collaboration at lower cost, and the two rewards add the main gains



Routing ≈ dense collaboration: 65.3 vs 65.4, without consulting experts on every question

+ CAR: 65.3 → 73.0; PMC-VQA multiple-choice 48.5 → 69.5

+ RR: → 74.7; open-ended questions benefit most (Slake 82.2 → 84.3)

Bar height is the mean accuracy over 7 settings (4 multiple-choice + 3 open-ended), computed from paper Table 4

Summary

1. **Collaboration should be selective:**
uncertainty-driven routing matches dense-collaboration accuracy with 1.17 calls per question and runs 3.8× faster

2. **Fully local deployment:**
a 4B model runs locally, so medical data never has to leave the premises

3. **Reasoning-centered training works:**
MedVR-3K plus dual-reward GRPO raises the 4-benchmark average by 6.7%

Code github.com/JayceonHo/SMART-Med
Data huggingface.co/datasets/JayceonHo/MedVR-3K
Paper arxiv.org/abs/2602.10179 Contact houjc24@mails.tsinghua.edu.cn

References

Bai et al. Qwen3-VL technical report. arXiv:2511.21631, 2025.

Chen et al. ReConcile: Round-table conference improves reasoning via consensus among diverse LLMs. ACL, 2024.

He et al. PathVQA: 30000+ questions for medical visual question answering. arXiv, 2020.

Kim et al. MDAgents: An adaptive collaboration of LLMs for medical decision-making. NeurIPS, 2024.

Lau et al. A dataset of clinically generated visual questions and answers about radiology images (VQA-RAD). Scientific Data, 2018.

Li et al. SMoA: Improving multi-agent LLMs with sparse mixture-of-agents. PAKDD, 2025.

Liu et al. SLAKE: A semantically-labeled knowledge-enhanced dataset for medical VQA. ISBI, 2021.

Shao et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models (GRPO). arXiv:2402.03300, 2024.

Tang et al. MedAgents: LLMs as collaborators for zero-shot medical reasoning. Findings of ACL, 2024.

Wang et al. Mixture-of-agents enhances large language model capabilities. ICLR, 2025.

Wang et al. ColaCare: Enhancing EHR modeling through LLM-driven multi-agent collaboration. WWW, 2025.

Xia et al. MMedAgent-RL: Optimizing multi-agent collaboration for multimodal medical reasoning. ICLR, 2026.

Zhang et al. Development of a large-scale medical visual question-answering dataset (PMC-VQA). Communications Medicine, 2024.

Zhu et al. MedAgentBoard: Benchmarking multi-agent collaboration with conventional methods. NeurIPS D&B, 2025.