MedEnv: Scaling Multimodal Virtual Medical Environments for Long-Horizon Diagnosis

Zhiting Fan[1*], Xuxiang Ren[2*], Yuan Wang[1], Ruizhe Chen[1], Zijie Meng[1], Keli Hu[3†], Jiaxiang Liu[1], Jian Wu[1,5], Weiqi Zhai[4], Hongxia Xu[1,5], and Zuozhu Liu[1,5†]

[1]Zhejiang University; [2]The University of Hong Kong; [3]Department of Computer Science and Engineering, Shaoxing University; [4]Alibaba Inc.; [5]Transvascular Implantation Devices Research Institute, Hangzhou, China
zhiting.23@intl.zju.edu.cn, zuozhuliu@intl.zju.edu.cn

Abstract. Real-world clinical care is a long-horizon decision-making process in which patient states and available evidence evolve over multiple turns, whereas static single-turn medical VQA is insufficient for training and evaluating interactive medical agents. To bridge this gap, we introduce **MedEnv**, a scalable multimodal virtual medical environment that supports multi-turn interviewing and tool-augmented evidence acquisition under realistic clinical workflows. Built from real patient records in MIMIC-IV, MedEnv transforms longitudinal EHR signals (admissions, laboratory tests, imaging, and clinical notes) into **496K**+ interactive trajectories for large-scale training and evaluation. Unlike prior EHR-only simulators, MedEnv integrates external disease-level medical databases for consistency-aware completion and constraint validation, reducing confounding from missing/noisy EHR signals and enabling longer interaction horizons. Leveraging this environment, we train a diagnostic agent with *confidence-based process supervision*, providing dense intermediate rewards tied to increases in confidence toward the correct diagnosis to reinforce informative questioning, retrieval, test ordering, and CXR analysis. Empirically, our trained agent **Diagnosis-R1** achieves 41.95% accuracy, surpassing strong baselines (best 37.56%). Scaling further improves performance but saturates around 50K RL trajectories, indicating that future gains likely require better data quality, hard-case coverage, and reward/sampling design. The code and data are available at https://github.com/FanZT6/MedEnv.

Keywords: Multimodal Diagnosis · Clinical Simulation · Medical Agent.

1 Introduction

Although medical large language models perform strongly on static evaluations (e.g., single-turn VQA) [3, 21, 22, 7], real-world clinical diagnosis more closely resembles a dynamic decision-making process: physicians iteratively refine differential diagnoses through multi-turn questioning and multimodal evidence such

[*]Equal contribution. [†]Corresponding author.

Table 1. Comparison with existing medical agents or benchmarks.

Model	Planning & Dialogue	Multi Modal	Tool Use	Text	Structured / EHR	Medical Image	Multi Turn	Num.
MMedAgent [11]	✓	✓	✓	✓		✓		48k
MedAgent [20]	✓			✓				8.9k
MedAgentBoard [23]	✓	✓	✓	✓	✓	✓		2.4k
MedAgent-Pro [19]	✓	✓	✓	✓	✓	✓		1.4k
DiagGym [14]	✓		✓	✓	✓		✓	2.2k
AgentHospital [12]	✓			✓	✓		✓	-
AI Hospital [4]	✓			✓	✓		✓	506
AgentClinic [16]	✓	✓	✓	✓	✓	✓	✓	1.6k
MedEnv (Ours)	✓	✓	✓	✓	✓	✓	✓	496k

as laboratory tests and imaging [6, 10, 18]. To support agents that can operate in this setting, we need an interactive environment: it should support turn-wise evolution of patient states, enable tool use that changes information availability, and provide feedback signals for long-horizon policy optimization.

Recent multi-turn diagnosis and tool-use benchmarks (Table 1) largely fall into EHR-driven and synthetic settings. EHR-based benchmarks better match clinical distributions but suffer from missing/noisy records, privacy constraints, and limited interaction logic and verifiability [4, 6, 11, 1, 16]. Synthetic benchmarks offer greater controllability, yet often deviate from real-world distributions and produce idealized or inconsistent evidence chains [18, 5].

To bridge this gap, we introduce **MedEnv**, a scalable **multimodal virtual medical environment** built from real patient records. MedEnv converts longitudinal EHR data (e.g., admissions, labs, clinical notes, radiology reports, and paired chest X-rays when available) into **496K+ interactive, multi-turn trajectories**, enabling large-scale training and evaluation under workflow-aligned interactions. A key challenge of EHR-driven simulation is the missingness and noise inherent to real-world records; MedEnv therefore augments patient-specific evidence with a **disease-level knowledge base** for consistency-aware completion and constraint-based validation, mitigating confounding signals and enabling longer interaction horizons. MedEnv further provides a workflow-aligned diagnostic tool suite (knowledge retrieval, imaging interpretation, structured access to tests/EHR fragments), enabling agents to strategically acquire and integrate evidence across sources.

Leveraging MedEnv, we train **Diagnosis-R1** with confidence-based process supervision to alleviate sparse terminal rewards in long-horizon diagnosis. Diagnosis-R1 uses a supervised cold start followed by reinforcement learning, optimizing a combined reward of (i) terminal diagnostic correctness and (ii) a dense **process reward** defined as the per-turn increase in log-probability of the ground-truth diagnosis after each action/tool call, encouraging informative evidence acquisition while penalizing low-yield interactions. Empirically, Diagnosis-R1 attains **41.95%** accuracy, surpassing strong baselines (best **37.56%**) while

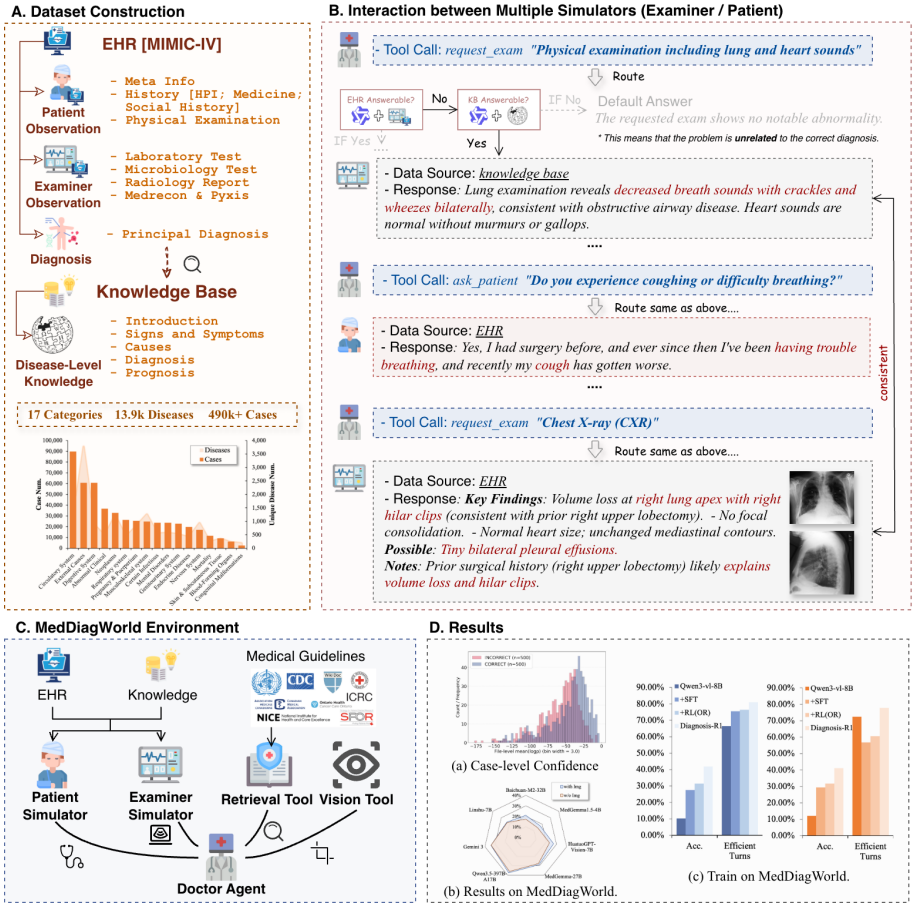

Fig. 1. Overview of MedEnv. (A) Dataset construction from real-world EHRs (MIMIC-IV) and a disease-level knowledge base. (B) Multi-turn, evidence-routed interaction between patient/examiner simulators with an EHR-first → knowledge-base augmentation → no-evidence fallback response mechanism. (C) The multimodal, tool-augmented environment. (D) Benchmark results and training gains on MedEnv.

staying interaction-efficient (**80.99%** efficient turns; **10.58** avg. turns). Imaging yields modest accuracy changes but more consistently improves efficiency by reducing low-yield steps. Further experiments confirm that scaling the amount of RL trajectories consistently improves performance, underscoring the value of our environment in providing better data and broader coverage.

2 Method

We construct a multi-role interactive environment for clinical consultation scenarios, centered on multi-turn dialogues among a patient simulator and a doctor

agent, for both training and evaluation on medical tasks, to simulate information gathering, evidence supplementation, and clinical decision-making in realistic consultations. Overall, the goal is to provide a training-and-testing platform that is realistic, controllable, and interpretable while remaining as faithful as possible to real clinical workflows.

2.1 Setups of MedEnv

Knowledge Base We build a disease-level knowledge base to complement missing EHR information. Specifically, we crawl and clean Wikipedia articles under `Diseases and Disorders by Systems`, yielding about 4.9k disease entries. At inference time, we match entries using ICD codes and disease cues extracted from the EHR, and retrieve the corresponding article when key evidence is missing to support more robust diagnostic reasoning.

Patient Simulator We implement an evidence-routing response mechanism for the patient simulator. Given a doctor's question, the environment first retrieves patient-specific evidence from the EHR and determines whether it is answerable from the record. If yes, it responds directly from the EHR to maintain factual consistency. Otherwise, it queries a disease-level knowledge base for supplementary information (e.g., symptoms, signs, or imaging patterns) and answers if relevant evidence is found; if not, it returns a standardized no-evidence fallback. Our EHR is constructed from real-world intensive-care records and paired chest radiographs, sourced from MIMIC-IV [8] and MIMIC-CXR [9], respectively. Overall, this "EHR-first → knowledge supplementation → no-evidence fallback" design supports natural multi-turn dialogue while preventing unsupported information from entering the doctor agent's decision-making.

Doctor Agent Settings During interaction, the doctor agent can conduct dialogue with the patient simulator to elicit symptoms, history, and other case-related information. When additional external evidence is needed, the agent may further invoke MCP tools as callable actions to retrieve or ground information, thereby better matching real clinical workflows.

Specifically, the callable tools include:

- **Retrieval Tool (`retrieve`)**: Retrieves guideline passages (e.g., diagnostic criteria, differential diagnosis, test recommendations, and management) based on symptoms or unresolved questions to support clinical reasoning.
- **Examination Tool (`image_exam`)**: Simulates ordering and reviewing imaging (e.g., CXR), returning the requested evidence (optionally with structured metadata) via verifiable tool calls.
- **Image Processing Tool (`grounding`)**: Performs visual grounding on imaging (via Grounding DINO [13]), returning candidate regions with confidence scores and visualizations to provide localized evidence (without producing diagnoses).

By explicitly modeling these tools as callable actions within the MCP interface, we enable more precise evaluation of the doctor agent's tool-use strategies, including when and how to invoke tools, the effectiveness of tool calls, and how external evidence integrates into subsequent decision-making processes.

2.2 Doctor Agent Training

We train the Doctor Agent using a two-stage pipeline: supervised fine-tuning (SFT) followed by reinforcement learning (RL). For SFT cold start, we synthesize ∼3k expert trajectories from Gemini-3-Pro, where each sample contains a complete interaction trajectory and the corresponding correct diagnosis.

Difficulty-aware data filtering. To stabilize RL and avoid reward signal collapse on trivial or unsalvageable cases, we perform a model-based filtering step using the SFT-initialized agent. For each candidate case, we run eight stochastic rollouts and retain only cases whose rollouts are *not* uniformly correct or uniformly incorrect. This keeps samples in a learnable regime (non-trivial yet solvable), improving reward variance and training stability.

RL optimization. We optimize the agent with RL using a reward composed of (i) an *outcome reward* r^{res} that scores the final diagnosis, and (ii) a dense *process reward* that assigns credit to intermediate interaction steps. The process reward is designed to encourage *information-seeking behavior* in long-horizon clinical interactions: the agent should ask questions or invoke tools only when they provide case-supported evidence that strengthens diagnostic belief, improving tool-use efficiency and reducing low-yield interactions.

Let y^* denote the ground-truth diagnosis for a case, and let $H_{<t}$ and $H_{\le t}$ be the interaction histories before and after turn t, respectively. We measure the *diagnostic information gain* at turn t by the change in log-probability assigned to the correct diagnosis, $g_t = \log \pi_\theta(y^* \mid H_{\le t}) - \log \pi_\theta(y^* \mid H_{<t})$, where $\pi_\theta(\cdot)$ is the policy model; larger g_t indicates more informative evidence acquisition (via dialogue or tool calls). We then define the process reward as $r_t^{\text{proc}} = g_t - \lambda_{\text{auto}} \mathbf{1}[a_t = \text{auto}]$, where the penalty discourages turns that trigger an automatic fallback (i.e., actions unsupported by EHR/KB evidence). The trajectory-level reward is $R = \lambda_{\text{res}} r^{\text{res}} + \sum_{t=1}^{T} r_t^{\text{proc}}$, with λ_{res} balancing terminal correctness and intermediate information gain over T interaction turns. We use terminal/process weights 1.0/0.5 and canonicalize targets with synonym, abbreviation, and ICD/disease-entry normalization.

3 Experiments

3.1 Experiment Setups

Data Processing We build a MedEnv data pipeline from MIMIC-IV, MIMIC-IV-Note, and optionally MIMIC-CXR, defining the pre-medication diagnostic phase as a unified observation window to avoid outcome leakage. We extract core EHR signals (demographics, admissions, labs, microbiology, medications,

Table 2. Comparison of model performance under settings w/o and with image.

Model	Without Image			With Image		
	Acc.	Eff. Turns	Avg. Turns	Acc.	Eff. Turns	Avg. Turns
General Model						
Qwen3-VL-8B [2]	10.28%	66.44%	11.027	12.16%	72.43%	9.14
Qwen3.5-397B-A17B [15]	35.99%	45.96%	11.15	37.56%	63.62%	10.76
GPT 5	34.84%	68.94%	13.47	35.86%	73.20%	11.37
Gemini-3-Flash	34.15%	65.80%	11.35	33.97%	78.52%	8.95
Medical Model						
MedGemma1.5-4B	17.49%	76.77%	12.11	14.59%	82.49%	11.75
HuluMed-4B [7]	18.62%	61.06%	12.94	20.87%	67.53%	15.42
Lingshu-7B [21]	22.69%	67.31%	19.65	23.17%	74.23%	19.87
HuatuoGPT-Vision-7B [22]	23.19%	74.73%	20.17	27.20%	80.12%	18.63
HuluMed-7b [7]	25.34%	71.06%	13.28	25.46%	72.36%	12.56
MedGemma-27B [17]	27.75%	48.29%	12.54	28.95%	46.33%	13.54
Baichuan-M2-32B [3]	17.34%	62.16%	9.6381	21.00%	68.84%	7.77
Hulu-Med-32B [7]	28.59%	70.28%	11.56	28.46%	70.08%	9.56
Diagnosis-R1	41.95%	80.99%	10.58	41.15%	77.87%	11.23

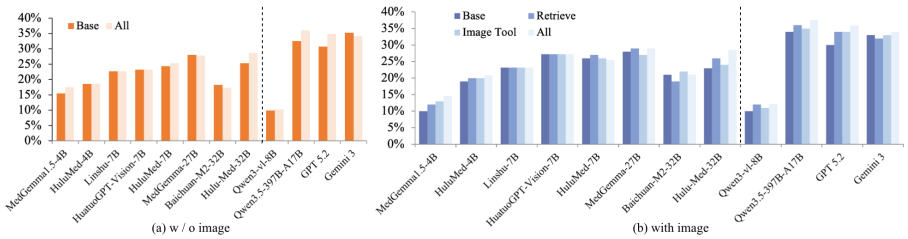

Fig. 2. Ablations on tool components across model scales.

orders/procedures, and service records) and parse history/physical exam notes, applying strict de-leakage to discharge summaries by keeping only early, pre-treatment content and removing sentences that explicitly mention the admission diagnosis. We also include radiology reports within the window and, when available, CXR images with paths and reports. Each admission is then packaged into a structured EHR episode (basic info, history, exams, labs, imaging, diagnosis labels) to form interactive trajectories for training and evaluation.

Evaluation We adopt an LLM-as-Judge protocol with **Qwen3-8B** to evaluate doctor-agent outcomes. We report **Accuracy** (whether the final diagnosis matches the ground-truth), and interaction efficiency via **Avg. Turns** (mean turns to termination) and **Efficient Turns**, which counts turns that yield verifiable, diagnostically useful evidence (e.g., informative patient replies or successful tool outputs), excluding low-yield steps such as redundant queries or failed/fallback tool calls. Together, these metrics capture the trade-off between diagnostic performance and interaction efficiency. The judge makes binary free-

Table 3. Ablation Study of Training Settings.

Model	Without Image			With Image		
	Acc.	Eff. Turns	Avg. Turns	Acc.	Eff. Turns	Avg. Turns
Base	10.28%	66.44%	11.027	12.16%	72.43%	9.14
+SFT	27.54%	75.46%	10.84	29.37%	56.86%	10.57
+RL (OR)	31.51%	76.37%	10.79	31.72%	60.63%	9.84
Diagnosis-R1	41.95%	80.99%	10.58	41.15%	77.87%	11.23

text diagnosis consistency decisions over synonyms/abbreviations; ICD matching serves as complementary validation.

Training Setup. We adopt Qwen3-VL-8B-Instruct as the base model and employ a two-stage pipeline: SFT followed by RL. We perform SFT with **ms-swift** on ∼3K diagnostic trajectories only as an interaction-format cold start. For RL, we initialize from the SFT checkpoint and optimize the policy with **GRPO** under **rllm** framework after difficulty-aware filtering. Training uses 8× NVIDIA H200 GPUs, group size 8, learning rate 1e-6, batch size 64, PPO mini-batch 32, and maximum horizon 15.

3.2 Main Results

We evaluate two model groups on the benchmark: domain-specific medical models and general-purpose models. Each model conducts multi-turn interactions with the environment to reach a diagnosis. Overall, adding images has model-dependent effects: it improves several baselines, but Diagnosis-R1 slightly decreases in accuracy and efficiency under the current CXR-only setting. These results suggest that visual evidence can lower interaction cost when integrated effectively, but may also introduce noisy or low-yield steps for agents already exploiting EHR and retrieval evidence strongly.

3.3 Ablation

We conduct ablation studies to assess the independent contributions of SFT, RL, and the process reward, all trained on top of the Qwen3-VL-8B-Instruct. Under a unified setting with the same model initialization, training data, and evaluation protocol, we construct: (1) a cold-start SFT-only model; (2) an SFT+RL variant without the process reward (i.e., using only the outcome reward); and (3) the full method (SFT+RL+process reward). All variants are trained with the same number of steps, sampling strategy, and hyperparameters, and are compared in terms of diagnostic accuracy, efficient turns, and average turns. The results are detailed in Table 3

The ablation results show that the `retrieve` component yields a clear improvement in diagnostic performance, with particularly pronounced gains for

Fig. 3. Scaling effects of RL training data on diagnostic accuracy.

smaller models (Fig. 2). This suggests that external medical knowledge retrieval effectively compensates for limited medical knowledge coverage and reasoning capacity in compact models, leading to better final diagnoses. In contrast, the impact of `grounding` is less significant. We hypothesize this is mainly because the current environment focuses on chest X-rays (CXR), where the most informative cues are often coarse-grained and do not require highly fine-grained visual detail, making the additional benefit of multimodal tools relatively limited in this setting.

3.4 Scaling of the Environment

We investigate how environment scale-the volume of RL training data-impacts model performance. Results show that diagnostic accuracy and interaction efficiency improve steadily with data size, as more trajectories provide richer learning signals. However, performance plateaus at approximately 50K samples, yielding diminishing returns. This suggests the model has reached the limits of current learning signals; further gains likely depend on data quality, hard-case coverage, and refined reward or training strategies rather than simple volume scaling.

4 Conclusion

We introduce **MedEnv**, a scalable virtual medical environment for long-horizon clinical diagnosis. In terms of environment construction, MedEnv emphasizes: (1) **comprehensiveness**—it is **multimodal** and exposes a workflow-aligned, extensible tool suite that can be readily expanded to new tools and evidence sources; (2) **multi-turn interaction**—it supports **multi-round** doctor–patient dialogue to evaluate and train agents' **active questioning** and evidence-seeking behaviors; and (3) **realism and reliability**—it is built upon real-world EHRs and incorporates **knowledge-base augmentation** to mitigate missing/noisy records while preserving verifiable, evidence-grounded responses. MedEnv further provides **496K**+ interactive trajectories, enabling large-scale training and

evaluation beyond the limited horizons and sizes of prior benchmarks. Leveraging MedEnv, we train **Diagnosis-R1** with confidence-based process supervision, demonstrating strong gains in diagnostic accuracy and interaction efficiency. MedEnv is a research environment rather than a deployable diagnostic system; EHR-first grounding and fallback policies reduce unsupported responses but do not establish clinical safety without clinician and external validation. We hope MedEnv will facilitate future research on reliable, tool-augmented medical agents, and plan to extend it with richer modalities and broader clinical workflows.

Acknowledgments. This work is supported by the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (Grant no. 2026C01021), Zhejiang Provincial Natural Science Foundation of China under Grant LZ24F020006, and Transvascular Implantation Devices Research Institute (TIDRI) under Grant No. KY052025008.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Almansoori, M., Kumar, K., Cholakkal, H.: Self-Evolving Multi-Agent Simulations for Realistic Clinical Interactions (2025). https://doi.org/10.48550/ARXIV.2503.22678
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Dou, C., Liu, C., Yang, F., Li, F., Jia, J., Chen, M., Ju, Q., Wang, S., Dang, S., Li, T., et al.: Baichuan-m2: Scaling medical capability with large verifier system. arXiv preprint arXiv:2509.02208 (2025)
4. Fan, Z., Wei, L., Tang, J., Chen, W., Siyuan, W., Wei, Z., Huang, F.: AI Hospital: Benchmarking Large Language Models in a Multi-agent Medical Interaction Simulator. arXiv preprint arXiv:2402.09742 (2024)
5. Gong, L., Fang, W., Yang, T., Tao, D., Guo, C., Wei, P., Xie, B., Guan, J., Chen, Z., Shi, F., Gu, J., Liu, J.: MedDialogRubrics: A Comprehensive Benchmark and Evaluation Framework for Multi-turn Medical Consultations in Large Language Models (Jan 2026). https://doi.org/10.48550/arXiv.2601.03023
6. Gong, L., Wang, A., Lai, Y., Ma, W., Liu, Y.: The Dialogue That Heals: A Comprehensive Evaluation of Doctor Agents' Inquiry Capability (2025). https://doi.org/10.48550/ARXIV.2509.24958
7. Jiang, S., Wang, Y., Song, S., Hu, T., Zhou, C., Pu, B., Zhang, Y., Yang, Z., Feng, Y., Zhou, J.T., et al.: Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. arXiv preprint arXiv:2510.08668 (2025)
8. Johnson, A.E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T.J., Hao, S., Moody, B., Gow, B., et al.: Mimic-iv, a freely accessible electronic health record dataset. Scientific data **10**(1), 1 (2023)
9. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019)

10. Lai, Y., Liu, K., Wang, Z., Ma, W., Liu, Y.: Doctor-R1: Mastering Clinical Inquiry with Experiential Agentic Reinforcement Learning (Feb 2026). https://doi.org/10.48550/arXiv.2510.04284
11. Li, B., Yan, T., Pan, Y., Luo, J., Ji, R., Ding, J., Xu, Z., Liu, S., Dong, H., Lin, Z., Wang, Y.: MMedAgent: Learning to Use Medical Tools with Multi-modal Agent. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 8745–8760. Association for Computational Linguistics, Miami, Florida, USA (2024). https://doi.org/10.18653/v1/2024.findings-emnlp.510
12. Li, J., Lai, Y., Li, W., Ren, J., Zhang, M., Kang, X., Wang, S., Li, P., Zhang, Y.Q., Ma, W., Liu, Y.: Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents (Jan 2025). https://doi.org/10.48550/arXiv.2405.02957
13. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: European conference on computer vision. pp. 38–55. Springer (2024)
14. Qiu, P., Wu, C., Liu, J., Zheng, Q., Liao, Y., Wang, H., Yue, Y., Fan, Q., Zhen, S., Wang, J., et al.: Evolving diagnostic agents in a virtual clinical environment. arXiv preprint arXiv:2510.24654 (2025)
15. Qwen Team: Qwen3.5: Towards native multimodal agents (February 2026), https://qwen.ai/blog?id=qwen3.5
16. Schmidgall, S., Ziaei, R., Harris, C., Reis, E., Jopling, J., Moor, M.: AgentClinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments (May 2025). https://doi.org/10.48550/arXiv.2405.07960
17. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
18. Tang, Y., Yu, J., Su, Z., Feng, K., Zhu, Z., Wang, L., Liang, L., Zhang, Q., Ding, K., Chen, H.: ClinDEF: A Dynamic Evaluation Framework for Large Language Models in Clinical Reasoning (Dec 2025). https://doi.org/10.48550/arXiv.2512.23440
19. Wang, Z., Wu, J., Cai, L., Low, C.H., Yang, X., Li, Q., Jin, Y.: MedAgent-Pro: Towards Evidence-based Multi-modal Medical Diagnosis via Reasoning Agentic Workflow (2025). https://doi.org/10.48550/ARXIV.2503.18968
20. Xu, R., Zhuang, Y., Zhong, Y., Yu, Y., Wang, Z., Tang, X., Wu, H., Wang, M.D., Ruan, P., Yang, D., Wang, T., Xiao, G., Liu, X., Yang, C., Xie, Y., Shi, W.: MedAgentGym: A Scalable Agentic Training Environment for Code-Centric Reasoning in Biomedical Data Science (2025). https://doi.org/10.48550/ARXIV.2506.04405
21. Xu, W., Chan, H.P., Li, L., Aljunied, M., Yuan, R., Wang, J., Xiao, C., Chen, G., Liu, C., Li, Z., et al.: Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. arXiv preprint arXiv:2506.07044 (2025)
22. Zhang, H., Chen, J., Jiang, F., Yu, F., Chen, Z., Chen, G., Li, J., Wu, X., Zhiyi, Z., Xiao, Q., et al.: Huatuogpt, towards taming language model to be a doctor. In: Findings of the association for computational linguistics: EMNLP 2023. pp. 10859–10885 (2023)
23. Zhu, Y., He, Z., Hu, H., Zheng, X., Zhang, X., Wang, Z., Gao, J., Ma, L., Yu, L.: MedAgentBoard: Benchmarking Multi-Agent Collaboration with Conventional Methods for Diverse Medical Tasks (2025). https://doi.org/10.48550/ARXIV.2505.12371