# OMNICLINIC: EVALUATING PROACTIVE EVIDENCE ACQUISITION IN A MULTIMODAL CLINICAL SANDBOX

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Medical artificial intelligence holds great promise for improving clinical care, making rigorous and realistic evaluation essential for its safe deployment. However, existing benchmarks assess clinical capability mainly through text-only dialogue, overlooking the proactive multimodal evidence acquisition required in real-world diagnosis. To address this gap, we introduce OMNICLINIC, an interactive multimodal sandbox and benchmark in which autonomous doctor agents acquire diagnostic evidence through clinical dialogue, diagnostic tests, and tool-mediated physical examinations. Agents operate within a finite clinical action space, selecting appropriate actions and specifying the parameters required for execution. We further introduce OMNIDOCTORX, which coordinates evidence-acquisition planning, spatial execution, and diagnostic review to support proactive clinical action. Across the evaluated omni-models, multimodal interaction improve diagnostic accuracy over text-only interaction. However, models struggle with evidence acquisition and spatial coordination, revealing a gap between multimodal understanding and autonomous interaction. OMNIDOCTORX improves mean proactive diagnostic accuracy from (86.36 vs. 77.84), demonstrating the value of coordinated acquisition and reasoning. Expert review further supports the data quality of the benchmark, strengthening the reliability as a sandbox for autonomous multimodal diagnosis.

## 1 INTRODUCTION

Despite rapid advances in digital health, in-person clinical visits remain the gold standard of healthcare delivery (Annaswamy et al., 2020). This necessity stems from a fundamental reality: medical assessment requires far more than verbal inquiry. Across the history of clinical practice, effective diagnosis has consistently relied on dynamic, multimodal perception and physical examination to uncover latent pathological signs (Balogh et al., 2015; McDuff et al., 2023). In Western medicine, this principle is codified in the canonical physical examination framework of IPPA (*Inspection, Palpation, Percussion, and Auscultation*), a philosophy mirrored in Traditional Chinese Medicine through the four diagnostic methods (*Wàng, Wén, Wèn, Qiè*: observation, auscultation and olfaction, interrogation, and palpation) (Kirby, 1981; Zhao et al., 2015). While large language models (LLMs) have demonstrated impressive proficiency in conversational triage and medical question answering, they remain largely confined to passive textual exchange (Wei et al., 2023; Gomez et al., 2024; Liu et al., 2023). Consequently, existing models cannot replicate the proactive, embodied, and multimodal interactions that offline physicians routinely deploy, creating a profound gap between conversational fluency and genuine clinical diagnostic competence (Balogh et al., 2015; Chiu et al., 2025).

Autonomous multimodal interaction remains underexplored in medical AI evaluation (Yang et al., 2026). Most benchmarks provide all relevant findings upfront and evaluate only diagnostic inference (Chiu et al., 2025; Yang et al., 2026). To address these evaluation gaps, we propose three principles for assessing autonomous multimodal diagnosis: 1) **Assessing Proactive Evidence Acquisition:** The agent should determine what evidence to seek and acquire it through clinical inquiry, physical examinations, and diagnostic tests, rather than receive all findings upfront. 2) **Assessing Coordinated Clinical Action Execution:** The agent should obtain evidence through valid actions, including selecting the appropriate body region and tool and configuring the pose and position of the patient and the doctor agent itself. 3) **Assessing End-to-End Diagnostic Effectiveness:** The benchmark should evaluate whether agents reach correct diagnoses through autonomous evidence acquisition and coordinated action in the clinical interactive settings.



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

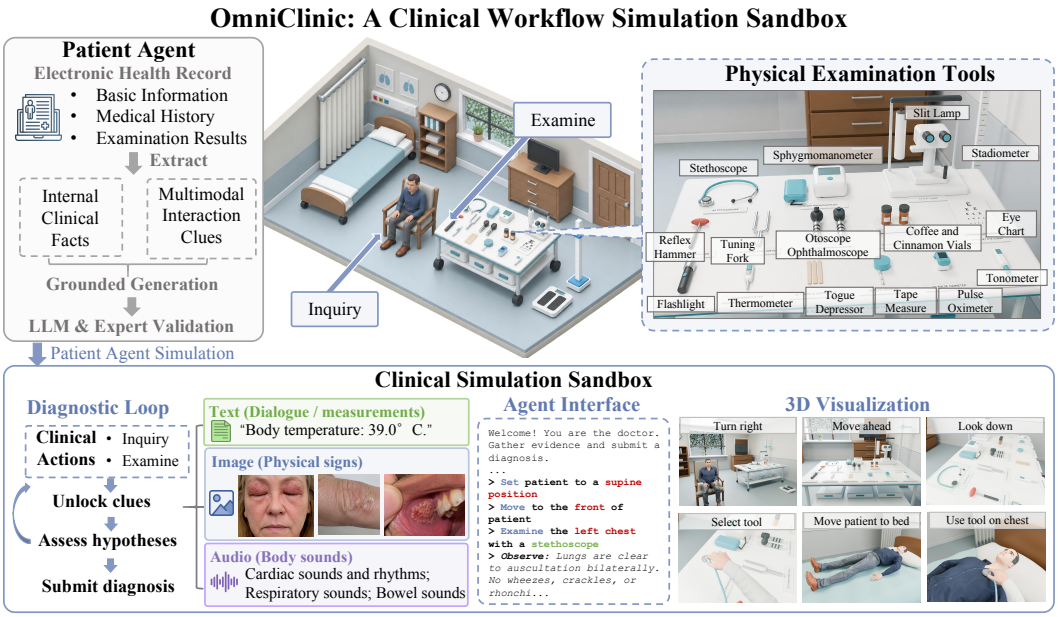


Figure 1: **Overview of OmniClinic.** Each electronic health record instantiates a patient agent and defines the case evidence. The sandbox exposes a finite action space spanning clinical consultation and bedside physical examinations with common clinical tools. The doctor agent selects among these actions to acquire multimodal evidence before submitting a diagnosis. Examination succeeds only with the appropriate tool, body region, patient pose, and doctor (agent) position.

Recent benchmarks have advanced interactive clinical evaluation through simulated patient consultation and sequential diagnosis (Schmidgall et al., 2026; Johri et al., 2025; Nori et al., 2025; Jiang et al., 2025; Liang et al., 2026). Typically, these environments expose findings through conversation or abstract tool calls to score the diagnosis (Chiu et al., 2025; Sviridov et al., 2025a). However, they do not jointly evaluate native multimodal acquisition, spatially grounded physical examination, and settings in which the evidence is either acquired by the model or supplied upfront (Chouayfati et al., 2026; Xu et al., 2026; Zhou et al., 2025). Consequently, high accuracy may reflect interpretation of available findings rather than the ability to identify and obtain the evidence needed for diagnosis. Furthermore, multimodal data of patient physiological states remain difficult to collect and annotate.

To this end, we introduce OMNICLINIC (Figure 1), an interactive multimodal clinical sandbox comprising 479 synthetic patient cases across 21 departments, with 4,251 images and 1,028 auscultation audio clips. Expert review supports generated data quality, while generation helps alleviate the scarcity of clinical data. Specifically, 1) To assess evidence acquisition, we compare **proactive interaction**, where agents acquire evidence through interaction, with **fully-presented evidence**, where the same reachable evidence is supplied upfront. 2) To assess action execution, the system checks examination requirements involving tools, body regions, patient poses, and doctor positions, and releases evidence only when the corresponding conditions are satisfied. 3) To assess diagnostic effectiveness, we evaluate final diagnostic accuracy alongside evidence acquisition and examination execution. We further introduce OMNIDOCTORX, an agentic framework that coordinates evidence-acquisition planning, spatial execution, and diagnostic review to improve autonomous clinical interaction.

We evaluate seven omni-model configurations on 479 cases and identify three key findings. 1) **Multimodal evidence improves diagnosis, but proactive interaction remains challenging.** All models achieve higher diagnostic accuracy with multimodal interaction than with text-only interaction, but providing evidence upfront yields higher accuracy. 2) **Spatial coordination and feedback-guided adaptation remain limited.** Only 1.68% of episodes include pose or position adjustments despite explicit instructions. After the first unsuccessful interaction, Qwen3-Omni-Thinking and Instruct repeat the same action in 88.24% and 65.51% of episodes respectively, highlighting difficulties in adapting acquisition strategies to unsuccessful interactions. 3) **Coordinated acquisition and reasoning improve diagnostic performance.** OMNIDOCTORX improves mean proactive diagnostic accuracy (86.36 vs. 77.84), while increasing evidence unlock rates, demonstrating the value of coordinating acquisition planning, spatial execution, and diagnostic review. Overall, OMNICLINIC

evaluates not only whether agents reach correct diagnoses, but also whether they can acquire evidence, coordinate examinations, and adapt their actions during clinical interaction.

## 2 Related work

Table 1: Comparison of modalities and interactive clinical capabilities across medical benchmarks.

| Benchmark | Modality | | Evidence Acquisition | | Physical Examination | | |
|---|---|---|---|---|---|---|---|
| | Image | Audio | Diagnostic Test | Proactive Evidence Acquisition | Physical Exam Tool Action | Body-site Targeting | Pose / Position Interaction |
| MedXpertQA (Zuo et al., 2025) | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| 3MDBench (Sviridov et al., 2025b) | ✓ | ✗ | ✗ | ✓ | ✗ | ✗ | ✗ |
| IMCBench (Xenochristou et al., 2026) | ✓ | ✗ | ✗ | ✓ | ✗ | ✗ | ✗ |
| AgentClinic (Schmidgall et al., 2026) | ✓ | ✗ | ✓ | ✓ | ✗ | ✗ | ✗ |
| SDBench (Nori et al., 2025) | ✗ | ✗ | ✓ | ✓ | ✗ | ✗ | ✗ |
| EviMed (Long & Wei, 2026) | ✗ | ✗ | ✓ | ✓ | ✗ | ✗ | ✗ |
| DDx-TRACE (Pan et al., 2026) | ✓ | ✗ | ✓ | ✓ | ✗ | ✗ | ✗ |
| MedAgentBench (Jiang et al., 2025) | ✗ | ✗ | ✓ | ✓ | ✗ | ✗ | ✗ |
| MedSP1000 (Liang et al., 2026) | ✗ | ✗ | ✓ | ✓ | ✗ | ✗ | ✗ |
| MedBench v5 (Ding et al., 2026) | ✓ | ✗ | ✗ | ✓ | ✗ | ✗ | ✗ |
| MedCTA (Ashraf et al., 2026) | ✓ | ✗ | ✗ | ✓ | ✗ | ✗ | ✗ |
| MedGaze-Bench (Liu et al., 2026) | ✓ | ✗ | ✗ | ✗ | ✗ | ✓ | ✗ |
| **OmniClinic (Ours)** | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

Table 1 highlights a shift from static multimodal evidence toward active clinical interaction. However, existing benchmarks typically evaluate only part of this process, while OmniClinic further requires agents to acquire evidence through physical examinations in an executable clinical environment.

**Medical Benchmarks** Medical evaluation has progressed from static evidence reasoning to interactive diagnosis. MedXpertQA evaluates fixed clinical questions with provided evidence (Zuo et al., 2025), while later benchmarks extend to image-grounded dialogue (Sviridov et al., 2025b; Xenochristou et al., 2026). Interactive benchmarks require models to acquire additional information through questions or examinations (Schmidgall et al., 2026; Nori et al., 2025; Long & Wei, 2026; Pan et al., 2026; Ding et al., 2026), some of them focus interaction to executable EHR tasks, standardized patient and software-tool-usage over medical images and reports (Jiang et al., 2025; Liang et al., 2026; Ashraf et al., 2026). Beyond information acquisition, recent work has also evaluated clinical spatial and temporal understanding through egocentric QA (Liu et al., 2026). Current benchmarks still lack a complete evaluation system of physical examination actions in executable clinical environments. OmniClinic instead evaluates sequential evidence acquisition in a clinical sandbox, connecting high-level action planning with semantic and spatial grounding for clinical interaction.

**Interactive Clinical Agents** Clinical agents have similarly progressed from primarily text-centric simulations to richer multimodal interaction. Earlier systems mainly model doctor-patient dialogue and the exchange of examination results (Li et al., 2025a; Kyung et al., 2026; Almansoori et al., 2025; Johri et al., 2025; Tu et al., 2025). Later work introduced more interactive patient behavior and spoken standardized-patient encounters (Voigt et al., 2025; Emerson et al., 2026; Johri et al., 2025), while sequential diagnostic agents began to choose follow-up questions and tests over multiple turns (He et al., 2026). Multimodal AMIE further incorporates visual evidence into diagnostic conversations (Saab et al., 2026). AMIE Video extends visual support into real-time audiovisual consultation and guides patients through remote physical examinations (Nagda et al., 2026). However, existing clinical agents remain largely conversational, with limited evaluation of whether they can translate diagnostic intent into grounded examination actions and adapt subsequent decisions to the observations they obtain. OmniClinic further turns physical examination into an executable interaction, where the agent must act in the environment to reveal clinical evidence.

## 3 OmniClinic

In this section, we present OmniClinic, an interactive clinical simulation sandbox with a diagnostic benchmark. The sandbox maintains interaction states, executes clinical actions, and releases evidence according to acquisition rules. We first formalize the interactive diagnosis task as a Partially Observable Markov Decision Process (POMDP) in Section 3.1, detail the multimodal dataset curation

pipeline, quality validation protocols, and summarize key benchmark statistics in Section 3.2, and then finally define our evaluation metrics in Section 3.3.

### 3.1 Problem Formulation: From Conversations to Multimodal Interactions

While clinical inquiry traditionally begins with verbal communication, accurate diagnosis fundamentally requires proactive multimodal interactions, incorporating bedside maneuvers and diagnostic investigations. We formalize this interactive clinical diagnostic task as a POMDP, defined by the tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \Omega, \mathcal{O}, T_{\max} \rangle$.

**State space.** The state $s_t = (z, x_t) \in \mathcal{S}$ consists of a static latent case $z \in \mathcal{Z}$ and a dynamic execution state $x_t \in \mathcal{X}$. The latent case contains the gold diagnosis and the complete inventory of history, physical findings, and test results. The dynamic state $x_t = (U_t, p_t, q_t, t, C_t)$ tracks unlocked evidence $U_t$, patient pose $p_t$, doctor position $q_t$, elapsed turns, and cumulative simulation cost. The agent observes information released from this state.

Table 2: Finite clinical action space.

| Function | Agent action | Environment return |
|---|---|---|
| Inquiry | Ask a target question | Patient report |
| Physical exam | Select tool + region | Multimodal finding |
| Auxiliary test | Order a named test | Test result |
| Spatial setup | Set agent pose / position | Updated state |
| Diagnosis | Submit a diagnosis | End episode |

**Action Space.** To eliminate operational ambiguity while retaining clinical flexibility, we retain natural language for verbal inquiry while operationalizing physical maneuvers and diagnostic tests as programmatic tool calls with typed parameters. The action space $\mathcal{A}$ comprises five functional classes:

- **Clinical Inquiry:** Asking free-form questions to obtain information about symptoms and history.
- **Physical Examination:** Selecting an examination method or tool and a body region to obtain clinical findings, subject to patient pose and doctor (agent) position requirements.
- **Diagnostic Testing:** Ordering auxiliary tests to obtain laboratory or imaging findings.
- **Spatial Configuration:** Adjusting the patient pose or the agent position to satisfy examination requirements, without directly revealing clinical evidence.
- **Diagnosis Submission:** Submitting the final diagnosis $\hat{z}$ and rationale to end the episode.

**Interaction Dynamics and Termination.** At turn $t$, the doctor agent selects an action $a_t \sim \pi(\cdot \mid h_t)$ based on the interaction history $h_t = (o_0, a_0, \ldots, a_{t-1}, o_t)$. The environment checks execution requirements and updates the state through $x_{t+1} = \mathcal{T}(x_t, a_t; z)$. The observation function $\mathcal{O}$ returns operational feedback and newly released findings in the corresponding modalities. Invalid calls return non-directive operational feedback. The episode ends when the agent submits a diagnosis or reaches the turn limit $T_{\max}$. The objective is to identify the correct diagnosis within this interaction horizon.

**Environment Implementation.** The clinical sandbox integrates patient simulation with state management, action execution, and rule-based evidence release. Inspired by Patient-Zero (Lai et al., 2026), patient simulation uses atomic facts to generate replies with consistency checks. Images and audio are released according to predefined acquisition rules. The doctor agent conducts inquiries, selects examinations and tools, and explicitly adjusts pose or position. Spatial states persist across turns, and multimodal evidence is released only when the action parameters satisfy the requirements. Unity serves only as a visualization prototype. Implementation details are listed in Appendix A.

### 3.2 Dataset Construction and Validation

OmniClinic builds on synthetic EHR cases generated by Patient-Zero (Lai et al., 2026), providing structured clinical records. These synthetic cases avoid exposing real patient data, while the structured generation and validation pipeline maintains clinical coherence and consistency. We extend each static record into a multimodal interactive case, in which actions such as questioning reveal patient basic information, taking physical examinations uncover physical findings, and ordering auxiliary tests return test results. To enable multimodal interaction, we instantiate suitable physical findings as image or audio observations that are revealed only through the corresponding clinical actions.

**Evidence Construction.** After filtering for valid identifiers, complete structured fields, and required assets, OMNICLINIC comprises 479 executable cases spanning 21 clinical departments. We organize each case into atomic evidence units: patient-reportable history is segmented into text units, auxiliary tests are remained, and physical-examination descriptions are decomposed into source-grounded findings. We then determine acquired action details and observation modality for each finding using predefined rules for common findings and model-assisted review for cases not covered by the rules. All annotations have source-consistency and completeness checks before remaining in the final dataset. The curation process includes iterative LLM-assisted review to refine validation rules.

**Multimodal Evidence Collection.** We use generative models with source-grounded prompts and one-shot references to instantiate visual and auditory findings. For images, prompts specify the atomic finding, polarity, anatomical region, examination context, and relevant demographics, while a canonical patient avatar provides a one-shot identity reference across images from the same case. Cardiac auscultation audio is generated with Stable Audio 3 Medium using the clinical finding and a licensed same-family recording as a one-shot acoustic reference, then standardized to six-second, 44.1-kHz mono waveforms. Detailed configuration and policy are listed in Appendix B.2.

**Benchmark Statistics.** The dataset contains 479 cases from 21 clinical departments. Each case specifies one primary diagnosis and hidden evidence inventory. The benchmark covers 295 distinct disease labels and 305 department-label pairs. Evidence inventory contains 45,953 text atoms and 5,279 native-media items. Images cover 7 clinical families and 41 concepts, while the current audio subset contains 1,028 auscultation clips. Complete data taxonomy is provided in Appendix B.3.

Table 3: Evidence inventory and coverage.

| Layer | Count | Scope |
|---|---|---|
| Cohort | 479 cases | 21 departments |
| Disease | 295 labels | 305 department-label pairs |
| Text | 45,953 atoms | History, physical findings, and test results |
| Image | 4,251 items | 7 families / 41 concepts |
| Audio | 1,028 items | Auscultation clips |

**Quality Validation and Audits.** Quality control combines automated checks, model-assisted media screening, and expert review. Automated checks verify data and media integrity, evidence accessibility, and the absence of diagnostic answers in public actions. Multimodal models assess image consistency with source findings and audio waveform and sound-type alignment. Final screening excludes 15.56% of candidate assets, and all 5,279 retained assets pass automated checks.

To assess the quality of the generated multimodal evidence, three physicians independently evaluated the image and audio sets in total 126 cases from 21 clinical departments using four criteria for each modality. For each case, the complete image and audio sets were rated separately on a five-point ordinal scale, yielding 1,008 complete criterion ratings. Image quality was assessed by *sign fidelity* (I1), *observability* (I2), *case consistency* (I3), and *media integrity* (I4); the corresponding audio criteria were *sound-sign fidelity* (A1), *audibility* (A2), *case consistency* (A3), and *media integrity* (A4). These criteria respectively measure accurate representation of the target finding, perceptual clarity, consistency with the case and other media, and freedom from artifacts or answer leakage.

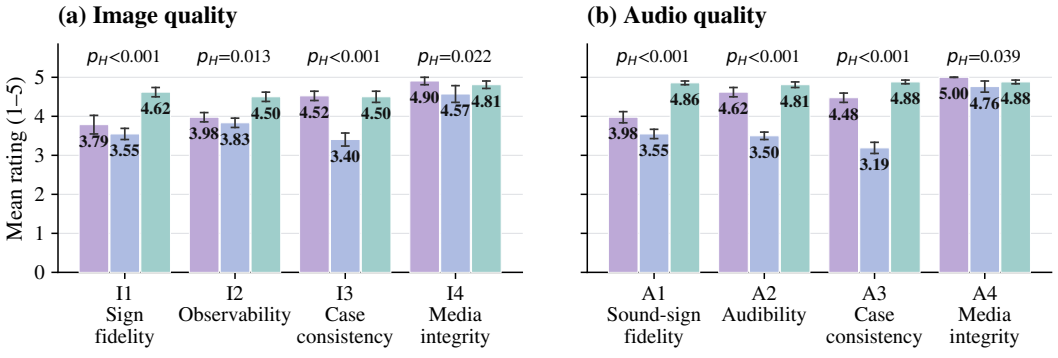


Figure 2: Physician-specific ratings of generated (a) images and (b) audio. Within each group, bars correspond from left to right to 3 experts. Error bars indicate pointwise 95% confidence intervals from bootstrap resampling of whole cases, stratified by department. $p_{\mathrm{H}}$ denotes permutation-Friedman $p$-values for rater differences, with Holm correction across the eight criteria.

As shown in Figure 2, the generated evidence received high overall ratings, with average scores of 4.25 for images and 4.38 for audio. Criterion-level scores ranged from 3.98 to 4.76 for images and from 4.13 to 4.88 for audio. Media integrity achieved the highest scores in both modalities, reaching 4.76 for images and 4.88 for audio. Inter-rater agreement was criterion dependent. Krippendorff's ordinal $\alpha$ ranged from $-0.220$ to $0.026$, indicating substantial differences in scoring behavior, whereas media integrity showed strong agreement under Gwet's AC2, with values of 0.867 for images and 0.937 for audio. Holm-adjusted permutation tests found significant rater effects for all eight criteria, although six remained significant in the department-level sensitivity analysis. Detailed metric definitions, statistical procedures, agreement analyses, and qualitative findings are provided in Appendix E.

### 3.3 Evaluation Metrics

We evaluate three aspects of agent performance: diagnostic accuracy, examination-argument accuracy, and media-unlock rates. Tool correctness and joint tool-region correctness measure the last outcome.

**1) Diagnostic accuracy**: Let $d_i$ be the gold diagnosis for episode $i$ and $\hat{d}_i$ the submitted diagnosis.

$$\mathrm{Acc} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left[\hat{d}_i = d_i\right]. \tag{1}$$

An episode is incorrect if the submitted label does not exactly match the gold label, if no diagnosis is submitted before the turn limit, or if the episode fails the evaluation protocol. Failure to submit within the turn limit and unrecoverable protocol errors are counted as incorrect.

**2) Examination-argument accuracy**: Let $\mathcal{Q}_i$ denote all recorded tool exam calls in episode $i$, including repeated and unsuccessful attempts but excluding proposals rejected before dispatch. For each call $q$, $I_t(q)$ indicates whether the tool matches a reachable examination route for the case, while $I_{tr}(q)$ requires the tool and body region to match the same route. We report

$$\mathrm{TC} = \frac{\sum_i \sum_{q\in\mathcal{Q}_i} I_t(q)}{\sum_i |\mathcal{Q}_i|}, \qquad \mathrm{TRC} = \frac{\sum_i \sum_{q\in\mathcal{Q}_i} I_{tr}(q)}{\sum_i |\mathcal{Q}_i|}. \tag{2}$$

Both metrics use the same call denominator, matching accounts for frozen public aliases and bilateral-region expansions. These scores measure compatibility with available case evidence, not general clinical validity or examination coverage. Rates are computed per seed and then averaged. Figure 6 extends this evaluation to combinations including patient pose and doctor (agent) position.

**3) Evidence-unlock rates**: For modality $m \in \{\mathrm{image}, \mathrm{audio}\}$, let $\mathcal{E}_m$ contain all evaluated episodes with at least one reachable item of that modality, and let $n_{i,m}$ be the number of such items acquired through clinical actions in episode $i$. We report

$$\mathrm{Unlock}_m = \frac{1}{|\mathcal{E}_m|}\sum_{i\in\mathcal{E}_m} \mathbb{I}[n_{i,m} > 0]. \tag{3}$$

Each episode contributes at most once, regardless of how many items are acquired. Image- and audio-unlock rates are computed separately within each seed and reported as percentages, with the mean and standard deviation across three seeds. These metrics measure whether agents access a modality, not acquisition completeness or interpretation accuracy.

## 4 How Can Doctor Agents Proactively Acquire Evidence?

Autonomous diagnosis requires coordinating what evidence to seek, how to obtain it through valid clinical actions, and how to revise diagnostic decisions as evidence accumulates. To address this coordination challenge, we introduce OmniDoctorX, an agentic framework that prioritizes **proactive interaction** of clinically relevant evidence. Using a fixed multimodal backbone, the doctor agent coordinates evidence acquisition, spatial execution, and independent diagnostic review based only on released observations. Implementation details are provided in Appendix C.

**Acquisition planning.** Given the observed history $h_t$, the backbone identifies unresolved findings and proposes relevant acquisitions from the public catalog, citing exposed evidence to justify their relevance. A deterministic compiler validates these proposals and constructs executable plans with admissible patient poses and doctor positions.

**Action selection.** Feasible candidates are ranked by $S(a \mid h_t) = D(a \mid h_t) + G(a \mid h_t)$, where $D$ is diagnostic utility gated by clinical relevance, and $G$ rewards relevant opportunities to acquire a missing native image or audio modality. This bonus combines relevance with estimated acquisition likelihood and vanishes once the modality is acquired. The controller executes the highest-scoring positive plan and replans from feedback. Evidence acquisition drives action selection, while costs serve as secondary criteria for spatial setup.

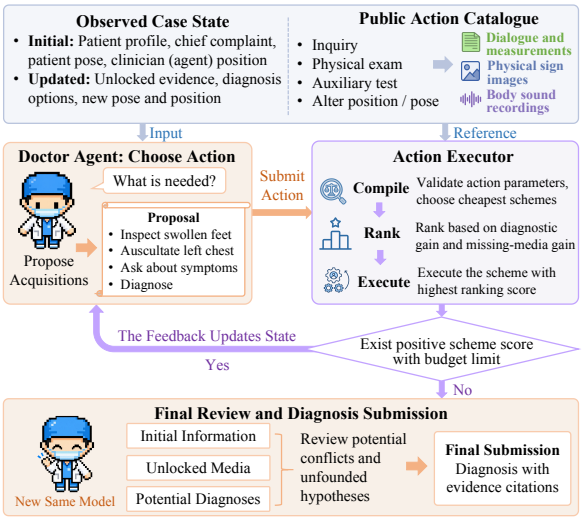


Figure 3: Overview of OMNIDOCTORX.

**Evidence-constrained diagnostic review.** Acquisition stops when no feasible proposed action has positive estimated value or the remaining action or model-call budget requires termination. A separate call to the same model then evaluates the original diagnosis options using only public initial information and acquired evidences. The reviewer does not see the doctor agent provisional diagnosis. Instead, it weighs supporting and conflicting evidence and flags unverified assumptions. Before submission, all evidence references are checked against the evidence released to the agent. The agent planning, stopping, and review rules are set before evaluation, while the benchmark action space, evidence-release, and scoring rules remain unchanged.

# 5 EXPERIMENTS

## 5.1 EXPERIMENTAL SETUP

**Evaluation protocol and models.** Evaluation covers 479 cases with three seeds. We compare **fully-presented evidence**, where evidence is supplied upfront, with **proactive interaction**, where agents acquire evidence through clinical actions. Proactive acquisition is evaluated with native Text+Image+Audio and text-only observations, with a limit of 15 turns per episode, including diagnosis submission. All experiments use the agent interface, while Unity serves only as a demonstration prototype. We evaluate seven multimodal model configurations: Gemini-3.7-Flash (Google DeepMind, 2026), the Instruct and Thinking variants of Qwen3-Omni-30B-A3B (Xu et al., 2025), MGM-Omni-32B (Wang et al., 2026), MiniCPM-o-4.5 (Cui et al., 2026), Phi-4-Multimodal-Instruct (Microsoft et al., 2025), and Baichuan-Omni-1.5 (Li et al., 2025b).

## 5.2 MAIN RESULTS

Table 4: Main results reported as mean$_{\pm\text{std}}$ over three seeds. Diagnostic accuracy compares **fully-presented evidence** (multimodal) with **proactive interaction** (multimodal and text-only). Action-execution and evidence-acquisition metrics are calculated under multimodal **proactive interaction** setting. Red values denote multimodal **proactive interaction** minus **fully-presented evidence**.

| Model | Diagnostic Accuracy | | | Action Execution | | Evidence Acquisition | |
|---|---|---|---|---|---|---|---|
| | Provided evidence | Proactive interaction (Multimodal) | Proactive interaction (Text-only) | Tool accuracy ↑ | Tool+body region accuracy ↑ | Image unlock ↑ | Audio unlock ↑ |
| Gemini-3.7-Flash | $90.07_{\pm 0.10}$ | $83.03_{\pm 0.23}$ (-7.03) | $71.43_{\pm 1.52}$ | $86.04_{\pm 3.03}$ | $79.61_{\pm 4.66}$ | $25.07_{\pm 1.26}$ | $3.31_{\pm 0.01}$ |
| Qwen3-Omni-Thinking | $85.46_{\pm 0.53}$ | $71.43_{\pm 0.79}$ (-14.03) | $63.88_{\pm 1.66}$ | $96.37_{\pm 1.28}$ | $94.87_{\pm 1.81}$ | $16.08_{\pm 0.36}$ | $10.30_{\pm 0.98}$ |
| Qwen3-Omni-Instruct | $85.39_{\pm 0.21}$ | $72.65_{\pm 0.91}$ (-12.74) | $65.83_{\pm 0.24}$ | $88.09_{\pm 0.35}$ | $77.40_{\pm 1.37}$ | $8.91_{\pm 0.12}$ | $28.38_{\pm 1.11}$ |
| MGM-Omni-32B | $80.32_{\pm 0.39}$ | $55.67_{\pm 0.43}$ (-24.64) | $53.68_{\pm 0.96}$ | $82.14_{\pm 2.32}$ | $56.69_{\pm 4.06}$ | $16.63_{\pm 0.79}$ | $4.40_{\pm 0.27}$ |
| MiniCPM-o-4.5 | $69.94_{\pm 1.10}$ | $66.67_{\pm 3.49}$ (-15.24) | $63.39_{\pm 2.32}$ | $91.64_{\pm 1.40}$ | $79.74_{\pm 1.21}$ | $29.02_{\pm 0.91}$ | $17.37_{\pm 0.59}$ |
| Phi-4-Multimodal-Instruct | $42.73_{\pm 1.42}$ | $38.07_{\pm 1.54}$ (-4.66) | $33.26_{\pm 1.54}$ | $99.28_{\pm 0.22}$ | $98.19_{\pm 0.80}$ | $32.08_{\pm 0.73}$ | $13.60_{\pm 0.36}$ |
| Baichuan-Omni-1.5 | $72.19_{\pm 0.72}$ | $62.26_{\pm 2.20}$ (-9.93) | $56.26_{\pm 3.32}$ | $39.53_{\pm 3.36}$ | $22.82_{\pm 3.21}$ | $52.05_{\pm 0.43}$ | $0.86_{\pm 0.27}$ |

Table 4 summarizes baseline performance on the 479-case benchmark, while Figure 4 compares the baselines with OMNIDOCTORX. Diagnostic accuracy covers **fully-presented evidence** (multimodal) and **proactive interaction** (multimodal and text-only). We highlight four findings.

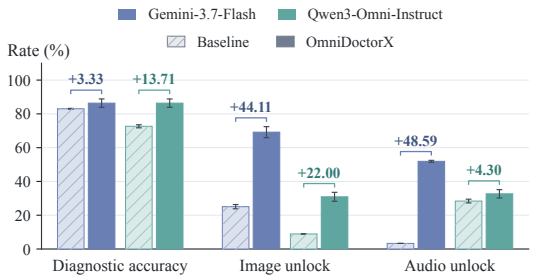


Figure 4: OMNIDOCTORX gains on diagnostic accuracy and media-unlock rates under multimodal **proactive interaction**. Error bars show standard deviations over three seeds. Annotations indicate absolute improvements in percentage points.

**1) Multimodal evidence improves diagnosis, but proactive interaction remains challenging.** As shown in Table 4, multimodal observations improve mean proactive interaction accuracy across all models. Multimodal observations improve diagnostic accuracy over text-only input for both Gemini-3.7-Flash (83.03 vs. 71.43) and Qwen3-Omni-Instruct (72.65 vs. 65.83). However, **diagnostic accuracy is consistently lower under proactive interaction than with evidence provided upfront**. Qwen3-Omni-Thinking (85.46 vs. 71.43 and Instruct (85.39 vs. 72.65) both achieve higher accuracy with **fully-presented evidence** than under **proactive interaction**. These comparisons highlight both the value of multimodal observations and the remaining gap between full evidence and autonomous clinical interaction.

**2) Anatomical grounding distinguishes examination-execution capabilities.** Table 4 reveals substantial differences in how reliably models associate tools with target body regions. MGM-Omni-32B shows 25.45% decrease from tool to tool-region accuracy (82.14 vs. 56.69). In contrast, Qwen3-Omni-Thinking largely preserves its performance (96.37 vs. 94.87). This contrast identifies anatomical grounding as a model-dependent execution bottleneck. Evaluating tool selection solely would overlook attempts that choose a compatible tool but fail to identify a corresponding body site.

**3) Native media remain unacquired in many diagnostic episodes.** The reported unlock rates imply that agents acquire image or audio evidence in no more than approximately 28.4% of episodes for Gemini and 37.3% for Qwen3-Omni-Instruct. Thus, a majority of episodes for both models unlock neither modality. This reveals a gap between multimodal evidence availability and actual acquisition. Diagnostic accuracy under a multimodal setting therefore reflects not only the interpretation of multimodal observations, but also whether those observations enter the diagnostic process.

**4) OMNIDOCTORX strengthens proactive interaction through coordinated planning, execution, and diagnostic review.** As shown in Figure 4, OMNIDOCTORX improves diagnostic accuracy for Gemini-3.7-Flash (86.36 vs. 83.03) and Qwen3-Omni-Instruct (86.36 vs. 72.65). Image and audio unlock rates increase by 44.11% and 48.59% for Gemini, and by 22.00% and 4.30% for Qwen. These gains are consistent with the framework ability to **convert diagnostic intent into executable evidence-acquisition plans**. Planning prioritizes relevant unresolved findings, spatial compilation coordinates patient pose and doctor position, and feedback guides subsequent acquisitions. Independent review then compares diagnoses against acquired evidence without seeing the provisional answer. Notably, Qwen3-Omni-Instruct achieves a larger diagnostic gain (13.71 vs. 3.33) despite smaller unlock gains, suggesting that evidence selection and interpretation matter alongside modality access.

### 5.3 ANALYSIS

This section analyzes spatial coordination and responses to failure feedback, followed by a case study comparing the baseline and OMNIDOCTORX, illustrating how different acquisition strategies shape the diagnostic process.

**Proactive spatial coordination remains limited despite explicit instructions.** The prompt explicitly instructs agents to check the patient pose and their own position before examinations and to adjust them when necessary. Nevertheless, only 169 of 10,054 eligible episodes (1.68%) across the seven models include a `move_patient` or `move_doctor` action. Figure 6 further shows substantial declines when patient pose is added to tool-region matching for Qwen3-Omni-Thinking (94.87 vs. 47.96) and Phi-4-Multimodal-Instruct (98.19 vs. 5.40). The combination of infrequent spatial actions and poor pose agreement suggests that existing omni models struggle with proactive spatial coordination despite explicit instructions.

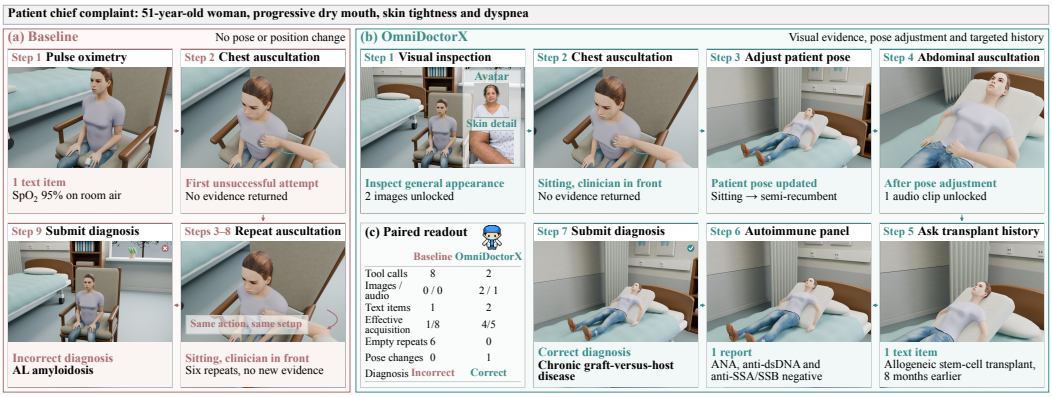


Figure 5: Case study comparing the baseline and OMNIDOCTORX with Qwen3-Omni-Instruct on the same case and seed. (a) Baseline trajectory, (b) OMNIDOCTORX trajectory, and (c) acquisition and diagnostic outcomes. The 3D scenes are illustrative constructions of actions and states.

**Does failure feedback lead to corrective action?** We examine the next decision after the first examination that returns no evidence in each episode, excluding cases without another acquisition opportunity. Pooling three seeds, Qwen3-Omni-Thinking and Instruct immediately repeat the same action, parameters, patient pose, and doctor position in 88.24% and 65.51% of these episodes, respectively. Such unchanged retries cannot reveal additional evidence in the static environment. Moreover, only 4.41% and 20.61% of their failure-conditioned episodes subsequently acquire new evidence through any channel. Gemini exhibits no immediate unchanged retries and subsequently obtains new evidence in 70.23% of its corresponding episodes. These patterns highlight **feedback-guided recovery as a distinct bottleneck in proactive diagnosis**. Interaction opportunities can be consumed by repeated unsuccessful attempts rather than broader exploration.

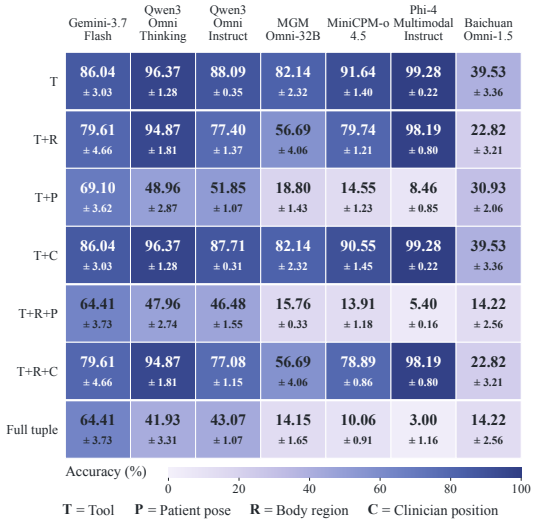


Figure 6: Joint accuracy of physical exam parameters under multimodal **proactive interaction**.

**Case study: adapting actions instead of repeating unsuccessful examinations.** Figure 5 compares acquisition strategies for the same backbone and patient case. The baseline obtains an oxygen-saturation measurement, then repeats an unsuccessful chest-auscultation call six times without changing its parameters or spatial configuration, before incorrectly predicting AL amyloidosis. In contrast, OMNIDOCTORX acquires two images, adjusts the patient's pose, and switches to abdominal auscultation to obtain an audio clip. It subsequently retrieves transplant history and an autoimmune panel before correctly identifying chronic graft-versus-host disease. Four of its five acquisition actions return new evidence despite fewer tool examinations compared to baseline (2 vs. 8). This selected example illustrates that **effective exploration requires revising the acquisition strategy when an examination yields no evidence**.

## 6 CONCLUSION

We introduced OMNICLINIC, an interactive multimodal benchmark for autonomous clinical diagnosis. Evaluation of seven omni-model configurations shows that multimodal evidence improves accuracy, but spatial coordination and feedback-guided adaptation remain limited. OMNIDOCTORX coordinates acquisition planning, spatial execution, and diagnostic review, improving mean proactive diagnostic accuracy from 77.84% to 86.36% across omni-models. Expert evaluation further supports data quality of the benchmark. Together, these contributions support systematic research on how medical agents acquire evidence, execute examinations, and reach diagnoses.

### AI USE STATEMENT

Beyond the LLM components used in the experiments, LLM assistance was used for manuscript drafting and revision, language editing, and checking consistency between the manuscript and supporting materials. Generative models were also used to create synthetic image and audio assets.

### ETHICS STATEMENT

This work investigates autonomous multimodal diagnosis in a controlled research environment. OMNICLINIC and OMNIDOCTORX are research prototypes and are not intended to provide medical advice, diagnosis, or treatment for real patients.

**Data Sources and Privacy** The benchmark builds on synthetic electronic health records from Patient-Zero (Lai et al., 2026), supplemented with generated images and auscultation audio. Reference recordings used during audio generation are distinct from the synthetic observations supplied to evaluated agents. Source datasets and reference assets remain subject to their applicable licenses and redistribution permissions.

**Human Evaluation Protocol** Three physicians independently evaluated the generated image and audio sets against the corresponding case descriptions and target findings. They used a five-point scale to assess finding fidelity, perceptual clarity, case consistency, and media integrity. This evaluation assessed dataset quality rather than model clinical skills or patient outcomes. Detailed procedures are provided in Section 3.2 and Appendix E.

**Annotator Welfare and Consent** Annotators were fully informed about the study's purpose, task design, and their right to withdraw at any time. All participation was voluntary, and compensation was provided at a fair rate consistent with local labor standards. No demographic or personal information about annotators was collected.

**Risks and Limitations** Generated evidence may contain clinical inaccuracies or demographic biases, and simulated examination requirements should not be interpreted as universal clinical standards. Automated screening and expert review do not eliminate these limitations. Performance on synthetic cases does not establish clinical safety or effectiveness, and generalization to real patient care requires separate validation.

### REPRODUCIBILITY STATEMENT

The evaluation protocol and model configurations are described in Section 5.1. The main evaluation uses a fixed cohort of 479 cases, three seeds, and a limit of 15 decision turns for interactive episodes. Action semantics, data construction, metric definitions, and agent prompts are documented in Appendices A, B, D, and F. The inference-time controller and diagnostic-review procedures of OMNIDOCTORX are detailed in Appendix C, with no updates to backbone model weights. Episode-level logs record actions, spatial states, released observations, and final predictions, supporting trajectory auditing and metric recomputation.

## REFERENCES

Mohammad Almansoori, Komal Kumar, and Hisham Cholakkal. MedAgentSim: Self-Evolving Multi-Agent Simulations for Realistic Clinical Interactions . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume LNCS 15968. Springer Nature Switzerland, September 2025.

Thiru M. Annaswamy, Monica Verduzco-Gutierrez, and Lex Frieden. Telemedicine barriers and challenges for persons with disabilities: Covid-19 and beyond. *Disability and Health Journal*, 13 (4):100973, 2020. doi: 10.1016/j.dhjo.2020.100973. URL https://doi.org/10.1016/j.dhjo.2020.100973.

Tajamul Ashraf, Hyewon Jeong, Fida Mohammad Thoker, and Bernard Ghanem. Medcta: A benchmark for clinical tool agents, 2026. URL https://arxiv.org/abs/2606.11702.

E. Peter Balogh, Bryan T. Miller, and John R. Ball (eds.). *Improving Diagnosis in Health Care*. National Academies Press, Washington, DC, 2015. doi: 10.17226/21794. URL https://doi.org/10.17226/21794.

Christopher Chiu, Silviu Pitis, and Mihaela van der Schaar. Simulating viva voce examinations to evaluate clinical reasoning in large language models. *arXiv preprint arXiv:2510.10278*, 2025. doi: 10.48550/arXiv.2510.10278. URL https://arxiv.org/abs/2510.10278.

Pia Chouayfati, Alexander M. Fichtl, Miriam Anschütz, George Doumat, and Georg Groh. Mtdiag: A multi-turn diagnostic dataset towards clinically meaningful llm evaluation. *arXiv preprint arXiv:2608.25085*, 2026. doi: 10.48550/arXiv.2608.25085. URL https://arxiv.org/abs/2608.25085.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, Jiancheng Gui, Luoyuan Zhang, Xian Sun, Fuwei Huang, Moye Chen, Zhuo Lin, Hanyu Liu, Qingxin Gui, Qingzhe Han, Yuyang Wen, Huiping Liu, Rongkang Wang, Yaqi Zhang, Hongliang Wei, Chi Chen, You Li, Kechen Fang, Jie Zhou, Yuxuan Li, Guoyang Zeng, Chaojun Xiao, Yankai Lin, Xu Han, Maosong Sun, Zhiyuan Liu, and Yuan Yao. Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction, 2026. URL https://arxiv.org/abs/2604.27393.

Jinru Ding, Chuchu Jiang, Lu Lu, Wenrao Pang, Mouxiao Bian, Zhuangzhi Gao, Jiangyuan Chen, Xinwei Peng, Ruiyao Chen, Sijie Ren, Renjie Lu, Yun Zhong, Bin Han, Meiling Liu, and Jie Xu. Medbench v5: A dynamic, process-oriented, and hallucination-aware benchmark for clinical multimodal models, 2026. URL https://arxiv.org/abs/2606.24155.

Andrew Emerson, Keelan Evanini, Su Somay, Kevin Frome, Le An Ha, and Polina Harik. Conversational AI for virtual standardized patients using a speech-to-speech LLM. In Giuseppe Riccardi, Seyed Mahed Mousavi, Maria Ines Torres, Koichiro Yoshino, Zoraida Callejas, Shammur Absar Chowdhury, Yun-Nung Chen, Frederic Bechet, Joakim Gustafson, Géraldine Damnati, Alex Papangelis, Luis Fernando D'Haro, John Mendonça, Raffaella Bernardi, Dilek Hakkani-Tur, Giuseppe "Pino" Di Fabbrizio, Tatsuya Kawahara, Firoj Alam, Gokhan Tur, and Michael Johnston (eds.), *Proceedings of the 16th International Workshop on Spoken Dialogue System Technology*, pp. 142–152, Trento, Italy, February 2026. Association for Computational Linguistics. URL https://aclanthology.org/2026.iwsds-1.16/.

Catalina Gomez, Junjie Yin, Chien-Ming Huang, and Mathias Unberath. How large language model-powered conversational agents influence decision making in domestic medical triage contexts. *Frontiers in Computer Science*, 6:1427463, 2024. doi: 10.3389/fcomp.2024.1427463. URL https://doi.org/10.3389/fcomp.2024.1427463.

Google DeepMind. Gemini 3.7 Flash: Model card. https://deepmind.google/models/model-cards/gemini-3-7-flash/, August 2026. Accessed: 2026-09-21.

Yufei He, Juncheng Liu, Zhiyuan Hu, Yulin Chen, Yue Liu, Yuan Sui, Yibo Li, Nuo Chen, Jun Hu, Bryan Hooi, Xinxing Xu, and Jiang Bian. Evoclinician: A self-evolving agent for multi-turn medical diagnosis via test-time evolutionary learning, 2026. URL https://arxiv.org/abs/2601.22964.

Yixing Jiang, Kameron C. Black, Gloria Geng, Danny Park, James Zou, Andrew Y. Ng, and Jonathan H. Chen. Medagentbench: A virtual ehr environment to benchmark medical llm agents. *NEJM AI*, 2(9):AIdbp2500144, 2025. doi: 10.1056/AIdbp2500144. URL https://ai.nejm.org/doi/full/10.1056/AIdbp2500144.

Shreya Johri, Jaehwan Jeong, Benjamin A. Tran, Daniel I. Schlessinger, Shannon Wongvibulsin, Leandra A. Barnes, Hong-Yu Zhou, Zhuo Ran Cai, Eliezer M. Van Allen, David Kim, Roxana Daneshjou, and Pranav Rajpurkar. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nature Medicine*, 31(1):77–86, 2025. ISSN 1546-170X. doi: 10.1038/s41591-024-03328-5. URL https://doi.org/10.1038/s41591-024-03328-5.

R. L. Kirby. Inspection-palpation-percussion-auscultation and an outcome-oriented alternative approach to the musculoskeletal examination. *Medical Education*, 15(2):106–109, 1981. doi:

10.1111/j.1365-2923.1981.tb02406.x. URL https://doi.org/10.1111/j.1365-2923.1981.tb02406.x.

Daeun Kyung, Hyunseung Chung, Seongsu Bae, Jiho Kim, Jae Ho Sohn, Taerim Kim, Soo Kyung Kim, and Edward Choi. Patientsim: A persona-driven simulator for realistic doctor-patient interactions. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026. URL https://openreview.net/forum?id=1THAjdP4QJ.

Yunghwei Lai, Ziyue Wang, Weizhi Ma, and Yang Liu. Patient-zero: Scaling synthetic patient agents to real-world distributions without real patient data, 2026. URL https://arxiv.org/abs/2509.11078.

Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, and Yang Liu. Agent hospital: A simulacrum of hospital with evolvable medical agents, 2025a. URL https://arxiv.org/abs/2405.02957.

Yadong Li, Jun Liu, Tao Zhang, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, Chong Li, Yuanbo Fang, Dongdong Kuang, Mingrui Wang, Chenglin Zhu, Youwei Zhang, Hongyu Guo, Fengyu Zhang, Yuran Wang, Bowen Ding, Wei Song, Xu Li, Yuqi Huo, Zheng Liang, Shusen Zhang, Xin Wu, Shuai Zhao, Linchu Xiong, Yozhen Wu, Jiahui Ye, Wenhao Lu, Bowen Li, Yan Zhang, Yaqi Zhou, Xin Chen, Lei Su, Hongda Zhang, Fuzhong Chen, Xuezhen Dong, Na Nie, Zhiying Wu, Bin Xiao, Ting Li, Shunya Dang, Ping Zhang, Yijia Sun, Jincheng Wu, Jinjie Yang, Xionghai Lin, Zhi Ma, Kegeng Wu, Jia li, Aiyuan Yang, Hui Liu, Jianqiang Zhang, Xiaoxi Chen, Guangwei Ai, Wentao Zhang, Yicong Chen, Xiaoqin Huang, Kun Li, Wenjing Luo, Yifei Duan, Lingling Zhu, Ran Xiao, Zhe Su, Jiani Pu, Dian Wang, Xu Jia, Tianyu Zhang, Mengyu Ai, Mang Wang, Yujing Qiao, Lei Zhang, Yanjun Shen, Fan Yang, Miao Zhen, Yijie Zhou, Mingyang Chen, Fei Li, Chenzheng Zhu, Keer Lu, Yaqi Zhao, Hao Liang, Youquan Li, Yanzhao Qin, Linzhuang Sun, Jianhua Xu, Haoze Sun, Mingan Lin, Zenan Zhou, and Weipeng Chen. Baichuan-omni-1.5 technical report, 2025b. URL https://arxiv.org/abs/2501.15368.

Cheng Liang, Pengcheng Qiu, Ya Zhang, Yanfeng Wang, Chaoyi Wu, and Weidi Xie. Evaluating large language models in dynamic clinical decision-making with standardized patient cases, 2026. URL https://arxiv.org/abs/2606.05112.

Shaonan Liu, Guo Yu, Xiaoling Luo, Shiyi Zheng, Jie Liu, Wenting Chen, and Linlin Shen. Benchmarking egocentric clinical intent understanding capability for medical multimodal large language models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26682–26697, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1228. URL https://aclanthology.org/2026.acl-long.1228/.

Xin Liu, Daniel McDuff, Geza Kovacs, Isaac Galatzer-Levy, Jacob Sunshine, Jiening Zhan, Ming-Zher Poh, Liao Shun, Paolo Di Achille, and Shwetak Patel. Large language models are few-shot health learners. *arXiv preprint arXiv:2305.15525*, 2023. doi: 10.48550/arXiv.2305.15525. URL https://arxiv.org/abs/2305.15525.

Zhuohan Long and Zhongyu Wei. Strong reasoning isn't enough: Evaluating evidence elicitation in interactive diagnosis. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 27559–27575, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1372. URL https://aclanthology.org/2026.findings-acl.1372/.

Daniel McDuff, Tao Schaekermann, Anil Tu, Amy Palepu, Anni Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, Le Cheng, Yun Liu, S. Sara Mahdavi, Sushrut Prakash, Anupam Pathak, Christopher Semturs, Ewa Dominowska, Joelle Gottweis, Jeanne Barral, Katherine Chou, Greg S. Corrado, Jake Sunshine, and V. Natarajan. Towards accurate differential diagnosis with large language models. *arXiv preprint arXiv:2312.00164*, 2023. URL https://arxiv.org/abs/2312.00164.

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras, 2025. URL https://arxiv.org/abs/2503.01743.

Mahvish Nagda, Jihyeon Lee, Matthew Thompson, Chunjong Park, Tim Strother, Valentin Liévin, Roma Ruparel, Akshay Goel, Teya Bergamaschi, Suhana Bedi, Meet Shah, Pavel Dubov, Liviu Panait, Toshiyuki Fukuzawa, Sam Schmidgall, Craig Schiff, Joseph Xu, Aliya Rysbek, Yana Lunts, Jan Freyberg, Rebecca Hemengway, Sunny Virmani, David Racz, Carey Radebaugh, Joëlle Barral, Kavi Goel, Dale R. Webster, Katherine Chou, Avinatan Hassidim, Yossi Matias, James Manyika, Gregory Wayne, Tao Tu, Yun Liu, Ethan Goh, Christina Chen, Ryutaro Tanno, Po-Hsuan Cameron Chen, Mike Schaekermann, and Anil Palepu. Towards expert-level medical ai for real-time video consultations, 2026. URL https://arxiv.org/abs/2608.09861.

Harsha Nori, Mayank Daswani, Christopher Kelly, Scott Lundberg, Marco Tulio Ribeiro, Marc Wilson, Xiaoxuan Liu, Viknesh Sounderajah, Jonathan Carlson, Matthew P Lungren, Bay Gross, Peter Hames, Mustafa Suleyman, Dominic King, and Eric Horvitz. Sequential diagnosis with language models, 2025. URL https://arxiv.org/abs/2506.22405.

Jiazhen Pan, Weixiang Shen, Jun Li, Julian Canisius, Felix Bitzer, Paula Roßmüller, Jiancheng Yang, Virginie Kreutzinger, Daniel Rueckert, and Benedikt Wiestler. Ddx-trace: A benchmark for medical diagnostic trajectories in vlms, 2026. URL https://arxiv.org/abs/2605.23629.

Khaled Saab, Chunjong Park, Tim Strother, Jan Freyberg, David G. T. Barrett, Yong Cheng, Wei-Hung Weng, David Stutz, Nenad Tomasev, Anil Palepu, Valentin Liévin, Yash Sharma, Roma Ruparel, Abdullah Ahmed, Elahe Vedadi, Kimberly Kanada, Cian Hughes, Yun Liu, Geoff Brown, Yang Gao, Sean Li, S. Sara Mahdavi, James Manyika, Katherine Chou, Yossi Matias, Avinatan Hassidim, Dale R. Webster, Joëlle Barral, S. M. Ali Eslami, Pushmeet Kohli, Adam Rodman, Vivek Natarajan, Mike Schaekermann, Tao Tu, Alan Karthikesalingam, and Ryutaro Tanno. Advancing conversational diagnostic ai with multimodal reasoning. *Nature Medicine*, 32(5):1726–1736, 2026. ISSN 1546-170X. doi: 10.1038/s41591-026-04371-0. URL https://doi.org/10.1038/s41591-026-04371-0.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Ji Woong Kim, Eduardo Pontes Reis, Jeffrey Jopling, and Michael Moor. Agentclinic: a multimodal benchmark for tool-using clinical ai agents. *npj Digital Medicine*, 9(1):499, 2026. ISSN 2398-6352. doi: 10.1038/s41746-026-02674-7. URL https://doi.org/10.1038/s41746-026-02674-7.

Ivan Sviridov, Amina Miftakhova, Artemiy Tereshchenko, Galina Zubkova, Pavel Blinov, and Andrey Savchenko. 3mdbench: Medical multimodal multi-agent dialogue benchmark. *arXiv preprint arXiv:2504.13861*, 2025a. doi: 10.48550/arXiv.2504.13861. URL https://arxiv.org/abs/2504.13861.

Ivan Sviridov, Amina Miftakhova, Artemiy Tereshchenko, Galina Zubkova, Pavel Blinov, and Andrey Savchenko. 3MDBench: Medical multimodal multi-agent dialogue benchmark. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26614–26654, Suzhou, China, November 2025b. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1353. URL https://aclanthology.org/2025.emnlp-main.1353/.

Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, Elahe Vedadi, Nenad Tomasev, Shekoofeh Azizi, Karan

Singhal, Le Hou, Albert Webson, Kavita Kulkarni, S. Sara Mahdavi, Christopher Semturs, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S. Corrado, Yossi Matias, Alan Karthikesalingam, and Vivek Natarajan. Towards conversational diagnostic artificial intelligence. *Nature*, 642 (8067):442–450, 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-08866-7. URL https://doi.org/10.1038/s41586-025-08866-7.

Henrik Voigt, Yurina Sugamiya, Kai Lawonn, Sina Zarrieß, and Atsuo Takanishi. Llm-powered virtual patient agents for interactive clinical skills training with automated feedback, 2025. URL https://arxiv.org/abs/2508.13943.

Chengyao Wang, Zhisheng Zhong, Bohao PENG, Senqiao Yang, Yuqi Liu, Haokun GUI, Bin Xia, Jingyao Li, Bei Yu, and Jiaya Jia. Mgm-omni: Scaling omni llms to personalized long-horizon speech. *ECCV*, 2026. URL https://www.paperdigest.org/?paper_id=eccv-4836-2026-09-10.

Qiuhong Wei, Zhengxiong Yao, Ying Cui, Bo Wei, Zhezhen Jin, and Ximing Xu. Evaluation of chatgpt-generated medical responses: A systematic review and meta-analysis. *arXiv preprint arXiv:2310.08410*, 2023. doi: 10.48550/arXiv.2310.08410. URL https://arxiv.org/abs/2310.08410.

Maria Xenochristou, Ashutosh Joshi, Korosh Vatanparvar, Mohammad Abuzar Hashemi, Prasad Kasu, Deepak Bansal, Anchal Nema, Nivedita Wadhwa, Prashams S Jain, Rebecca Abraham, Will Kimbrough, Dilek Hakkani-Tur, and Wilko Schulz-Mahlendorf. Imcbench: A benchmark for multimodal llms in image-grounded medical conversations, 2026. URL https://arxiv.org/abs/2606.28556.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report, 2025. URL https://arxiv.org/abs/2509.17765.

Shihao Xu, Tiancheng Zhou, Jiatong Ma, Yanli Ding, Yiming Yan, Ming Xiao, Guoyi Li, Haiyang Geng, Yunyun Han, Jianhua Chen, and Yafeng Deng. Lingxidiagbench: A multi-agent framework for benchmarking llms in chinese psychiatric consultation and diagnosis. *arXiv preprint arXiv:2602.09379*, 2026. doi: 10.48550/arXiv.2602.09379. URL https://arxiv.org/abs/2602.09379.

Rui Yang, Weihao Xuan, Lin Yi, Jonathan Chong Kai Liew, Matthew Yu Heng Wong, Nicolás Lescano, Nikita R. Paripati, Mona Diab, Hua Xu, David W. Bates, Yifan Peng, et al. Evaluating multi-turn multimodal diagnostic reasoning on challenging real-world clinical cases. *arXiv preprint arXiv:2607.25933*, 2026. doi: 10.48550/arXiv.2607.25933. URL https://arxiv.org/abs/2607.25933.

Changbo Zhao, Guo-Zheng Li, Chengjun Wang, and Jinling Niu. Advances in patient classification for traditional chinese medicine: A machine learning perspective. *Evidence-Based Complementary and Alternative Medicine*, 2015:376716, 2015. doi: 10.1155/2015/376716. URL https://doi.org/10.1155/2015/376716.

Tianhong Zhou, Yin Xu, Yingtao Zhu, Chuxi Xiao, Haiyang Bian, Lei Wei, and Xuegong Zhang. Drvd-bench: Do vision-language models reason like human doctors in medical image diagnosis? *Advances in Neural Information Processing Systems*, 38, 2025. URL https://arxiv.org/abs/2505.24173.

Yuxin Zuo, Shang Qu, Yifei Li, Zhang-Ren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. MedxpertQA: Benchmarking expert-level medical reasoning and understanding. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=IyVcxU0RKI.

# A Action Execution and Evidence Release

## A.1 Public Actions and Spatial State

Agents interact with a deterministic clinical environment through a case-independent action vocabulary. The public context includes the initial patient profile, chief complaint, diagnosis options, action schemas, and examination catalog. Unreleased findings, reference labels, source records, and case-specific acquisition routes remain server-side. All actions are executed by `SandboxSession`. The Unity client serves only as a demonstration prototype.

Table 5 lists the required parameters, possible observations, and original action costs used during evaluation. Patient pose and doctor position are persistent environment states. They change only through explicit spatial actions and are not inferred from an examination request. The doctor is the doctor agent itself.

Table 5: Public actions, required parameters, possible observations, and abstract simulation costs.

| Action | Required parameters | Returned information | Cost |
|---|---|---|---|
| *Clinical inquiry* | | | |
| `ask` | `question` | Patient report (text) | 1 |
| *Spatial configuration* | | | |
| `move_patient` | `pose` | Updated patient pose, no evidence | 1 |
| `move_doctor` | `position` | Updated doctor position, no evidence | 1 |
| *Physical examination* | | | |
| `visual_exam` | `body_region` | Text or image findings | 2 |
| `palpation_exam` | `body_region` | Text or image findings | 2 |
| `percussion_exam` | `body_region` | Text findings | 2 |
| `tool_exam` | `tool`, `body_region` | Text, image, or audio findings | 3 |
| `maneuver_exam` | `exam_type`, `body_region` | Text or image findings | 3 |
| `cognitive_exam` | `task` | Text findings | 3 |
| *Auxiliary investigation* | | | |
| `test` | `test_name` | Recorded test report (text) | 6 |
| *Diagnostic submission* | | | |
| `submit_diagnosis` | `diagnosis` | Submission confirmation and termination | 1 |

*Notes.* Only schema-required parameters are listed. Observation types refer to the native multimodal condition. They indicate possible returns rather than guaranteed evidence.

## A.2 Evidence Release, Feedback, and Termination

**Clinical inquiry.** Patient responses use deterministic retrieval over patient-reportable evidence rather than an additional patient language model. Each question returns at most one matching text atom with its evidence ID. Unmatched questions receive an uncertainty response. Repeating a question does not rotate through undisclosed findings.

**Examinations and investigations.** An examination releases recorded evidence only when its parameters and current spatial state match an accepted acquisition route. Public tool-region compatibility indicates procedural feasibility, not the presence of a finding in the current case. Auxiliary tests return available recorded reports. The environment does not synthesize normal findings for unavailable results.

**Observation delivery.** Released findings carry evidence IDs that agents may cite in subsequent decisions. Images and audio are delivered as native attachments in the multimodal condition. Text-only evaluation removes image and audio observations without supplying their textual descriptions. Spatial adjustments update the environment state but reveal no clinical evidence.

**Unsuccessful actions.** An examination without matching evidence returns non-directive feedback. This feedback does not identify the correct hidden route or distinguish a setup mismatch from an

unavailable finding. An unavailable result therefore does not establish a normal clinical finding. Cases are static, so repeating an examination with unchanged parameters and spatial state does not reveal additional evidence.

**Costs and termination.** Every executed action incurs its original cost from Table 5, including repetitions and unsuccessful examinations. Rejected decisions consume a decision turn but release no evidence and incur no action cost. Episodes end upon diagnosis submission or after 15 decision turns, with diagnosis submission required on the final turn. Accumulated simulation cost does not terminate an episode. Cost aggregation and supplementary results are reported in Appendix D.5.

**Diagnosis submission.** The agent selects an exact diagnosis option and may cite only exposed evidence IDs. Submission ends the episode and is scored against the hidden reference label. Reference answers and correctness feedback are not available during evidence acquisition or diagnostic review.

# B Data Construction and Dataset Characteristics

## B.1 Case Construction and Evidence Grounding

**Frozen cohort.** The benchmark contains 479 synthetic patient cases across 21 clinical departments, with 4,251 images and 1,028 auscultation recordings. Each case has a unique patient identifier, a source-record link, and an evidence inventory. Dataset preparation checks identifiers, schemas, asset integrity, and acquisition reachability before model evaluation. The prepared environment contains 45,953 text atoms. This total includes text derived from patient information, history, and auxiliary reports in addition to the 7,899 text clues in the original multimodal annotations. The distribution tables in Appendix B.3 describe the original annotated clue inventory, rather than the complete prepared text inventory.

**Evidence atomization.** Patient-reportable evidence is extracted from complaint and history fields, excluding objective measurements, examination findings, diagnoses, and treatment instructions. Each nonempty auxiliary-examination entry becomes a separately orderable text atom. Physical findings are extracted with gpt-4o-mini in JSON mode at temperature zero. Extraction preserves polarity, laterality, severity, timing, and measurements, splitting findings when they require different acquisition actions. A deterministic sentence fallback retains content missed by the model and flags it for review.

**Modality and action mapping.** A versioned registry maps each finding to its observation modality, acquisition type, public action, and required parameters. Sounds map to audio, visible findings to images, and measurements or manual-examination outcomes to text. Observation modality describes the returned information rather than the underlying procedure. For example, a prose radiology report is a text observation. Tools and body regions are normalized against a case-independent public catalog. Runtime routes may include bilateral expansions and alternative pose-position combinations. These alternatives are operational simulation rules rather than unique clinical examination standards. Original-annotation matching is evaluated separately in Appendix D.

**Validation and information boundaries.** Preparation assigns stable evidence IDs, normalizes aliases, removes answer-bearing fields from public action parameters, and validates actions against the public schema. Acquisition routes are replayed in `SandboxSession` to check whether legal actions under attainable conditions release the intended evidence. All 5,279 media assets pass the mechanical integrity checks. Source records, reference labels, and case-specific acquisition routes remain hidden from the evaluated agent. Automated validation establishes structural integrity and operational consistency, not comprehensive clinical validity. Expert assessment of sampled image and audio sets is reported separately in Appendix E.

## B.2 Multimodal Generation and Quality Control

**1) Target extraction.** Each media target corresponds to an atomic finding and retains its patient identifier, evidence identifier, polarity, anatomical region, acquisition method, and source provenance. Duplicate targets are removed. Audio targets are assigned to a sound family, such as heart, lung, bowel, or vascular sounds. Unsupported targets or those lacking an appropriate reference are withheld from generation.

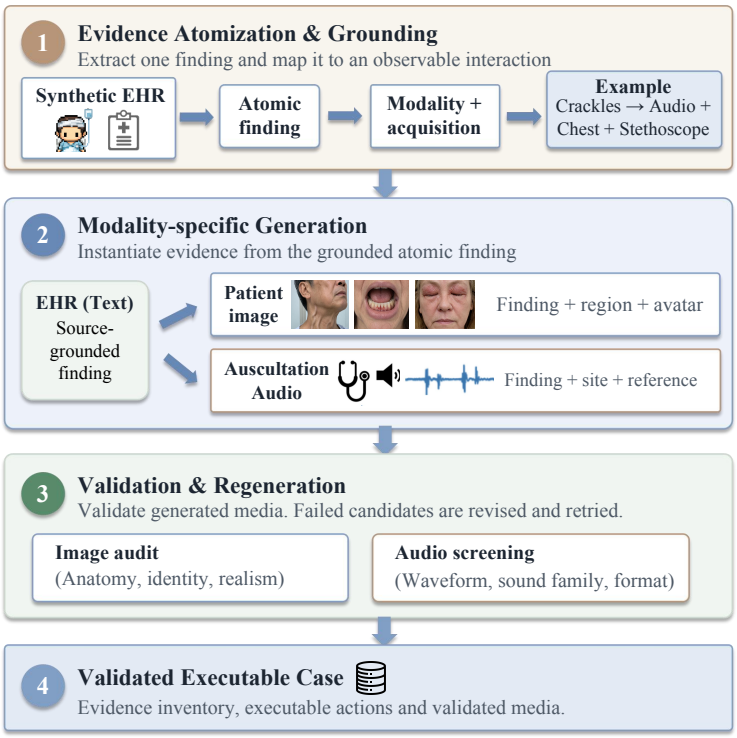


Figure 7: Multimodal data construction through evidence grounding, modality-specific generation, and iterative validation, producing quality-controlled executable patient cases.

**2) Image generation and identity consistency.** Image prompts specify the target finding, polarity, body region, acquisition method, and limited demographic attributes. Supporting context is retrieved from the source record, retaining at most four deduplicated fragments ranked by overlap with the finding and region. Prompts preserve negation and laterality, distinguish clinical photographs from medical imaging, and prohibit unsupported abnormalities, collages, annotations, and multiple views. The diagnosis may serve as server-side background context during initial generation, but neither it nor the generation prompt is exposed to the evaluated agent.

Images are generated with gpt-image-2 at $1024 \times 1024$ resolution. Before generation, gpt-4o-mini at temperature zero checks clinical realism, source consistency, polarity, anatomy, laterality, and imaging modality. Failed prompts are revised and rechecked for up to five rounds. Prompt lengths are capped at 3,800 characters for detail images and 5,000 for avatars. Each patient receives a canonical avatar, preferably an existing avatar or one derived from the earliest general-appearance image. The earliest available patient image is used as a fallback. Detail images preserve the avatar's appearance and clothing while retaining their clinical findings. Avatars are not themselves treated as diagnostic evidence.

**3) Audio generation and screening.** Audio prompts specify the finding, polarity, auscultation site, sound family, and available physiological rate. They exclude speech, music, alarms, electronic tones, and unrelated sounds. Each target uses a reference recording from the corresponding sound family, preferring a matching finding and a normal reference for negated abnormalities. Reference recordings support generation and screening, while the released audio is synthetic.

Audio is generated with Stable Audio 3 Medium using audio-and-text conditioning without fine-tuning. Candidates use a six-second duration, eight diffusion steps, classifier-free guidance scale $1.0$, and deterministic per-attempt seeds. Initial reference-noise levels are $0.12$ for negated findings, $0.16$ for normal findings, and $0.32$ for other positive findings. Retries vary this level within $[0.08, 0.55]$. Outputs are converted to mono, 16-bit PCM WAV at $44.1$ kHz. Structural checks reject decoding failures, incorrect formats, duration deviations exceeding $0.15$ seconds, near-silence, and clipping.

The AuscultaBase encoder provides a coarse sound-family check. Candidates and references are converted to mono 16-kHz audio, silence-trimmed, padded or truncated to eight seconds, and

represented by 64-bin log-Mel spectrograms. Let $\phi(y)$ and $\phi(r_c)$ denote normalized embeddings of candidate $y$ and the reference for family $c$. A candidate passes only when

$$c = \arg\max_{c'} \phi(y)^\top \phi(r_{c'}), \qquad \phi(y)^\top \phi(r_c) \geq 0.35, \tag{4}$$

and the similarity to the expected family exceeds every competing family by at least $0.02$. This check assesses coarse sound-family consistency rather than clinical correctness of every acoustic finding.

**4) Validation and regeneration.** Generated images undergo multimodal audits for finding accuracy, anatomical localization, acquisition realism, identity consistency, composition, and artifacts. All required criteria must pass. Failed candidates are regenerated using audit feedback, with at most five image-generation attempts or ten audio-generation attempts. Targets that remain unresolved are withheld rather than silently accepted. Of 2,041 initial audio targets, 1,028 recordings across 424 cases are retained after regeneration and capability filtering. The remaining 1,013 targets are withheld as unsupported or unresolved. The released audio subset is restricted to cardiac sounds. These filters are applied before doctor-agent evaluation. Expert review provides a separate assessment of data quality rather than a replacement for these automated checks.

### B.3 Dataset Composition and Coverage

**Clinical coverage.** The 479 cases span 21 departments and 295 globally distinct diagnosis labels. The distribution is long-tailed, with 210 labels appearing in only one case. Department-specific counts sum to 305 because some labels occur in multiple departments. The labels also vary in specificity, including 62 department-specific ancestor-descendant relationships and two synonym pairs. Table 6 shows the detailed clinical coverage, with examples list up to three frequent diagnosis labels per department.

Table 6: Clinical coverage of the 479 cases.

| Department | Cases | # Diagnoses | Example diagnoses |
|---|---|---|---|
| Cardiology | 19 | 13 | Acute Decompensated Heart Failure (5), Heart Failure with Preserved Ejection Fraction (2), Ischemic Heart Disease (2) |
| Dentistry | 13 | 7 | Oral and Maxillofacial Space Infection (4), Ludwig's Angina (2), Oropharyngeal Squamous Cell Carcinoma (2) |
| Dermatology | 50 | 19 | Contact Urticaria (7), Cutaneous Tuberculosis (7), Acne (5) |
| Emergency | 40 | 29 | Airway Obstruction (4), Disseminated Intravascular Coagulation (3), Acute Heart Failure (2) |
| Endocrinology | 17 | 13 | Graves' Disease (2), Hyperthyroidism (2), Post-transplantation Diabetes Mellitus (2) |
| Gastroenterology | 6 | 6 | Acute Non-Variceal Upper Gastrointestinal Bleeding (1), Autoimmune Gastritis (1), Celiac Disease (1) |
| General Surgery | 19 | 16 | Diabetic Foot (3), Esophageal and Gastric Variceal Bleeding (2), Severe Acute Pancreatitis (2) |
| Hematology | 29 | 23 | Hypoproliferative Thrombocytopenia (3), Cryoglobulinemia (2), Disseminated Intravascular Coagulation (2) |
| Immunology | 15 | 10 | Gout (3), Anti-MDA5 Antibody-Positive Dermatomyositis (2), Malignant Hypertension (2) |
| Infectious | 16 | 16 | Biliary Tract Infection (1), Brucella Endocarditis (1), Carbapenem-Resistant Gram-Negative Bacterial Infection (1) |
| Nephrology | 35 | 13 | Chronic Kidney Disease-Mineral and Bone Disorder (CKD-MBD) (10), Renal Anemia (7), Sarcopenia (7) |
| Neurology | 49 | 17 | Benign Paroxysmal Positional Vertigo (6), Fatal Familial Insomnia (6), Menstrual Migraine (6) |
| Obstetrics and Gynecology | 17 | 10 | Acute Fatty Liver of Pregnancy (5), Preeclampsia (3), Chronic Hypertension with Superimposed Preeclampsia (2) |
| Oncology | 34 | 27 | Brain Parenchymal Metastases (3), Non-Small Cell Lung Cancer Brain Parenchymal Metastases (3), Cancer-Related Anorexia (2) |
| Ophthalmology | 15 | 14 | Macular Degeneration (2), Acute Angle-Closure Glaucoma (1), Adenoid Cystic Carcinoma of the Lacrimal Gland (1) |

Continued on the next page

Table 6 continued

| Department | Cases | # Diagnoses | Example diagnoses |
|---|---|---|---|
| Orthopedics | 6 | 5 | Post-traumatic Limb Edema (2), Femoral Fracture (1), Open Injury (1) |
| Otolaryngology | 21 | 15 | Deep Neck Abscess (3), Hearing Impairment (3), Benign Paroxysmal Positional Vertigo (2) |
| Pediatrics | 12 | 8 | Dravet Syndrome (3), Developmental Coordination Disorder (2), Developmental Language Disorder (2) |
| Psychiatry | 41 | 26 | Obsessive-Compulsive Personality Disorder (5), Gaming Disorder (3), Parkinson's Disease Dementia (3) |
| Respiratory | 13 | 11 | Highly Pathogenic Avian Influenza in Humans (2), Usual Interstitial Pneumonia (2), Acute Exacerbation of Chronic Obstructive Pulmonary Disease (1) |
| Urology | 12 | 7 | Prostate Cancer with Neuroendocrine Differentiation (5), Primary Bilateral Macronodular Adrenocortical Disease (2), Adrenocortical Carcinoma (1) |
| **Total** | **479** | **305** | **295 globally unique diagnosis labels** |

**Patient demographics.** Adults and older adults account for more than 95% of cases, while the sex distribution is relatively balanced. Profiles cover eight ethnicity categories, six continental regions, 44 nationalities, and 96 residential cities. Table 7 summarizes demographic, socioeconomic, and lifestyle attributes.

Table 7: Demographic and social composition of the 479 cases.

| Attribute | Category | % |
|---|---|---|
| Age | Adult (18 - 64 years) | 67.01 |
| | Older adult ($\geq$65 years) | 28.18 |
| | Child ($<$18 years) | 4.59 |
| | Neonate ($<$28 days) | 0.21 |
| Sex | Male | 53.44 |
| | Female | 46.56 |
| Ethnicity | Asian | 36.74 |
| | Caucasian | 25.89 |
| | African | 19.00 |
| | Hispanic | 12.11 |
| | Middle Eastern | 3.13 |
| | Native American | 1.46 |
| | Mixed | 1.25 |
| | Pacific Islander | 0.42 |
| Residence | Urban | 64.09 |
| | Rural | 19.42 |
| | Suburban | 16.49 |
| Smoking | Never | 60.96 |
| | Current | 21.09 |
| | Former | 17.95 |
| Alcohol use | None | 57.62 |
| | Moderate | 32.99 |
| | Heavy | 9.39 |

| Attribute | Category | % |
|---|---|---|
| Geographic region | Asia | 39.25 |
| | Africa | 19.62 |
| | North America | 18.37 |
| | Europe | 13.36 |
| | South America | 5.43 |
| | Oceania | 3.97 |
| Employment status | Employed | 82.88 |
| | Retired | 12.32 |
| | Student | 2.30 |
| | No stated occupation | 2.51 |
| Socioeconomic status | Middle income | 62.42 |
| | Low income | 23.38 |
| | High income | 9.39 |
| | No income | 4.80 |
| Lifestyle | Moderate | 49.27 |
| | Sedentary | 38.83 |
| | Active | 11.90 |
| Diet | Balanced | 58.66 |
| | High-fat/high-salt | 30.90 |
| | Vegetarian | 10.23 |
| | Other | 0.21 |
| Physiological state | Normal | 92.28 |
| | Postmenopausal | 5.01 |
| | Pregnant | 2.71 |

**Annotated evidence inventory.** The original multimodal annotations contain 13,178 clues, comprising 7,899 text, 4,251 image, and 1,028 audio items. These counts exclude the additional text atoms introduced during environment preparation. Text and image clues appear in all 21 departments. Audio is available in 424 cases, while 55 cases contain no audio evidence.

**Acquisition types.** Observation modality specifies what information is returned, whereas acquisition type specifies how it is obtained. Table 9 summarizes acquisition annotations for the same 13,178 clues. Tool-based examinations cover 15 tools and can return text, image, or audio evidence. These annotation-level counts are distinct from the number of actions executed by an agent.

Table 8: Department-level distribution of the original annotated clues.

| Department | Text | Image | Audio |
|---|---|---|---|
| Cardiology | 251 | 107 | 71 |
| Dentistry | 196 | 115 | 11 |
| Dermatology | 629 | 593 | 91 |
| Emergency | 643 | 317 | 80 |
| Endocrinology | 306 | 180 | 55 |
| Gastroenterology | 116 | 75 | 3 |
| General Surgery | 267 | 145 | 17 |
| Hematology | 475 | 240 | 54 |
| Immunology | 215 | 123 | 37 |
| Infectious | 283 | 134 | 38 |
| Nephrology | 596 | 330 | 132 |
| Neurology | 912 | 379 | 100 |
| Obstetrics and Gynecology | 297 | 122 | 37 |
| Oncology | 567 | 289 | 79 |
| Ophthalmology | 325 | 138 | 18 |
| Orthopedics | 87 | 48 | 5 |
| Otolaryngology | 418 | 221 | 11 |
| Pediatrics | 215 | 137 | 26 |
| Psychiatry | 822 | 378 | 116 |
| Respiratory | 102 | 85 | 23 |
| Urology | 177 | 95 | 24 |
| **Total** | **7,899** | **4,251** | **1,028** |

Table 9: Acquisition types associated with the original annotated clues.

| Acquisition type | Description | Count | % |
|---|---|---|---|
| Visual inspection | Inspect the patient or a visible physical finding. | 4,260 | 32.33 |
| Tool use | Use an examination tool: stethoscope, sphygmomanometer, thermometer, pulse oximeter, weight scale, stadiometer, tape measure, flashlight, otoscope-ophthalmoscope, pocket eye chart, slit lamp, tonometer, tongue depressor, reflex hammer, or 128-Hz tuning fork. | 3,618 | 27.45 |
| Palpation | Examine a body region by touch or applied pressure. | 2,588 | 19.64 |
| Physical maneuver | Perform a maneuver and observe the resulting finding or response. | 1,764 | 13.39 |
| Questioning | Ask for symptoms, history, or other patient-reportable information. | 948 | 7.19 |
| **Total** | | **13,178** | **100.00** |

# C OmniDoctorX Implementation Details

OmniDoctorX uses the same public observations, action vocabulary, and environment as the baseline. It introduces an inference-time controller for acquisition selection, spatial execution, and independent diagnostic review. The controller does not access hidden findings or patient-specific acquisition routes, and no backbone weights are updated.

## C.1 Candidate Generation and Selection

**Candidate generation.** At each planning step, the backbone proposes up to two candidates in each of three groups: image acquisition, audio acquisition, and other acquisition actions. Groups may be empty. Each candidate specifies a public action, its parameters, a target finding, and exposed evidence IDs supporting its clinical relevance. The backbone also provides a provisional diagnosis, which does not veto a positive acquisition candidate.

**Candidate scoring.** Each candidate $a$ receives ordinal ratings for relevance $r_a \in \{0, 1, 2\}$, acquisition likelihood $\ell_a \in \{0, 1, 2, 3\}$, and diagnostic value $v_a \in \{0, 1, 2, 3\}$. Let $M(a, h_t)$ indicate that

the candidate targets an image or audio modality for which no native evidence has yet been acquired. The controller uses

$$\begin{aligned} D(a \mid h_t) &= \mathbb{I}[r_a > 0]\frac{v_a}{3}, \\ G(a \mid h_t) &= 2M(a, h_t)\frac{r_a}{2}\frac{\ell_a}{3}, \\ S(a \mid h_t) &= D(a \mid h_t) + G(a \mid h_t). \end{aligned} \tag{5}$$

These ratings are ordinal judgments rather than calibrated probabilities. The media bonus becomes zero once the target modality has actually been acquired. Proposed examinations, textual findings, and predicted outcomes do not count as native-media acquisition. The controller selects the highest-scoring feasible candidate with strictly positive utility. Ties favor greater missing-modality gain, fewer primitive actions, lower original action cost, and finally a stable public semantic key. Simulation cost and model-call cost are not penalties in the primary acquisition score.

### C.2 Spatial Execution and Stopping Rules

**Spatial compilation.** A deterministic compiler validates action parameters and identifies admissible examination setups from the public catalog. It inserts the required patient or doctor movements before the selected examination. Every movement and examination remains an original environment action and counts toward the 15-turn limit. Public setup compatibility does not guarantee that the patient has a corresponding finding.

**Repetition and replanning.** Previously dispatched examination selectors are excluded from automatic repetition. Cosmetic changes to target descriptions or argument wording do not create a new examination opportunity. After execution, the controller updates its evidence inventory using actual feedback and replans. No particular examination, modality, or fixed sequence is mandatory.

**Stopping.** Each acquisition plan must leave one primitive action available for diagnosis submission. Acquisition ends when no positive feasible candidate remains or when the action or model-call allowance requires termination. The provisional diagnosis is then passed to the submission workflow only after the independent review described below.

### C.3 Diagnostic Review and Inference Budgets

**Independent diagnostic review.** A separate call to the same backbone receives the public initial information, acquired observations, attached native media, and original diagnosis options. The acquisition policy's provisional diagnosis, candidate scores, and proposed-answer rationale are withheld. Earlier invalid proposals are sanitized to avoid exposing their proposed answers. The reviewer compares up to three diagnosis options using supporting evidence, conflicting evidence, and unresolved assumptions. Missing observations remain unknown rather than being treated as normal findings. The reviewer selects one exact diagnosis option and cites exposed evidence IDs. Identifier validation checks whether evidence was exposed, not whether its clinical interpretation is correct.

**Budgets and repairs.** Acquisition, diagnostic review, and format repair share an allowance of 45 completed model calls. Up to three calls are reserved for review, allowing an initial response and at most two repair attempts. Repairs append `previous_interface_error` without supplying additional clinical evidence. Exhausted review repairs produce a model failure. Spatial setup and diagnosis submission remain within the unchanged 15-primitive interaction limit. Complete agent instructions and output schemas are provided in Appendix F.

## D Evaluation Metrics and Additional Results

### D.1 Episode Accounting and Result Aggregation

**Episode accounting.** Diagnostic accuracy includes all eligible case-seed episodes, including model failures, rather than only successful submissions. Submitted diagnoses are scored against the reference labels. Episodes ending without a valid submission, including turn-limit termination and unrecovered model protocol errors, count as incorrect. Unresolved provider or transport failures are recorded separately and excluded from model-performance denominators.

**Result aggregation.** For the main results and joint examination-argument accuracy, each metric is computed separately within each seed. We report the mean and sample standard deviation across three seeds. Metrics with a zero denominator are reported as unavailable rather than zero. The failure-feedback rates described below instead pool episode counts across seeds.

**Media acquisition.** Image- and audio-unlock rates use eligible episodes containing at least one reachable item of the corresponding modality. An episode counts as successful if it unlocks at least one such item, regardless of subsequent acquisitions. These rates measure modality access rather than evidence completeness or interpretation accuracy.

**Spatial adjustment.** An episode contributes to the spatial-adjustment rate if its recorded trajectory contains at least one `move_patient` or `move_doctor` action. The aggregate rate in Section 5.3 pools eligible episodes across the seven models and three seeds. This episode-level measure is distinct from the call-level annotation agreement defined below.

### D.2 Joint Physical Examination Parameters Accuracy

**Evaluation scope.** We evaluate recorded `tool_exam` calls using four components: tool ($T$), body region ($R$), patient pose ($P$), and doctor position ($C$). Tool and region are taken from the examination parameters, while pose and position are taken from the environment state at execution. The evaluated component sets are

$$\mathcal{S} = \{T, TR, TP, TC, TRP, TRC, TRPC\}.$$

The doctor position is the position of the doctor agent itself.

**Original-annotation matching.** Let $\mathcal{R}_i^0$ denote the original annotation routes for reachable evidence in case $i$. For an evaluated component set $S$, a call $q$ is correct only if all components in $S$ jointly match one route:

$$J_S(q) = \mathbb{I}\left[\exists r \in \mathcal{R}_i^0 \text{ such that } \bigwedge_{k \in S} \mathrm{match}_k(q, r)\right]. \tag{6}$$

Components cannot be matched independently against different routes. Unlisted components are ignored. Matching retains the frozen tool and region aliases and bilateral-region expansions, but excludes spatial alternatives added by the runtime. Full-tuple correctness requires simultaneous agreement on all four components.

**Denominator and aggregation.** Let $\mathcal{Q}_{i,s}$ contain all recorded `tool_exam` calls for eligible episode $i$ under seed $s$. Repeated calls and attempts yielding no evidence remain in the denominator. Proposals rejected before dispatch as `tool_exam` are excluded. For every component set,

$$\mathrm{Acc}_s(S) = 100 \frac{\sum_i \sum_{q \in \mathcal{Q}_{i,s}} J_S(q)}{\sum_i |\mathcal{Q}_{i,s}|}. \tag{7}$$

This is a call-level rate, not an average of episode-level percentages. All component sets use the same denominator within each seed. A zero denominator is reported as unavailable rather than zero.

For $K = 3$ seeds, we report

$$\begin{aligned} \overline{\mathrm{Acc}}(S) &= \frac{1}{K} \sum_{s=1}^{K} \mathrm{Acc}_s(S), \\ \mathrm{SD}(S) &= \sqrt{\frac{1}{K-1} \sum_{s=1}^{K} \left(\mathrm{Acc}_s(S) - \overline{\mathrm{Acc}}(S)\right)^2}. \end{aligned} \tag{8}$$

**Consistency and interpretation.** For nested component sets $S \subseteq S'$, joint matching guarantees

$$\mathrm{Acc}_s(S') \leq \mathrm{Acc}_s(S).$$

The same ordering holds for seed-averaged accuracy, and equality is allowed. It does not impose an ordering on non-nested combinations such as $TR$ and $TP$, or on their standard deviations.

Table 10: Matched modality ablation for Gemini-3.7-Flash. Diagnostic accuracy (%) is reported as mean ± standard deviation across three seeds.

| Condition | Available observations | Diagnostic accuracy |
|---|---|---|
| T | Text | $71.43 \pm 1.52$ |
| TI | Text+Image | $80.71 \pm 0.66$ |
| TA | Text+Audio | $76.88 \pm 0.67$ |
| TIA | Text+Image+Audio | $83.03 \pm 0.23$ |

These scores measure agreement with original annotations rather than general clinical validity, examination coverage, or runtime evidence-unlocking success. They are computed post hoc without rerunning the interactions under stricter rules. Since models choose different examinations and repeated calls are counted, comparisons also reflect attempted-action distributions. Matching a complete route does not by itself establish that the agent actively changed its pose or position rather than satisfying the route from the default state.

### D.3 Modality Ablation and Case Study

**Matched evaluation.** We examine Gemini-3.7-Flash under four observation conditions: Text (T), Text+Image (TI), Text+Audio (TA), and Text+Image+Audio (TIA). All conditions use the same frozen case data, execution code, action space, and 15-turn limit, with matching diagnosis-option orders across three seeds. These recorded runs use the benchmark protocol with case-grounded retrieval for textual responses. All conditions retain access to textual history, examination findings, and test reports. Withheld media are not replaced by textual descriptions.

**Aggregate results.** Table 10 shows that both visual and auditory observations improve mean diagnostic accuracy over text-only input, with gains of 9.28 and 5.45 percentage points, respectively. Combining both modalities achieves the highest accuracy (83.03%), exceeding TI and TA by 2.32 and 6.15 points. These results suggest complementary benefits from visual and auditory evidence, with images contributing a larger improvement in this setting. Since agents select their own actions in each condition, these comparisons capture the combined effects of evidence acquisition and interpretation rather than isolating perception under a fixed trajectory.

The case study below highlights that modality ablation in interactive diagnosis evaluates both whether agents acquire the available evidence and how they use it. Accuracy differences therefore cannot be attributed to multimodal interpretation alone.

**Case study: image acquisition and diagnostic decisions**

**Case and pairing.** A synthetic 65-year-old man presents with progressive left upper abdominal discomfort, abdominal fullness, and unintentional weight loss. The reference diagnosis is adrenocortical carcinoma. We compare the same case (`Urology_line425`) and seed across all four conditions.

| Condition | Images / audio acquired | Final prediction |
|---|---|---|
| T | 0 / 0 | Renal cell carcinoma (incorrect) |
| TI | 2 / 0 | Adrenocortical carcinoma (correct) |
| TA | 0 / 0 | Renal cell carcinoma (incorrect) |
| TIA | 2 / 0 | Adrenocortical carcinoma (correct) |

**Observed trajectories.** TI and TIA perform abdominal visual inspection at steps 7 and 9, respectively, obtaining the same two images. Their final rationales explicitly cite these images and refer to abdominal striae and distension before selecting the reference diagnosis. T and TA acquire no native media and select renal cell carcinoma.

**Interpretation and scope.** This example documents explicit use of acquired images in diagnostic reasoning, but does not isolate their causal contribution because the trajectories and acquired textual evidence also differ. No audio is acquired in any condition, so the case cannot establish an audio benefit or image-audio synergy.

Table 11: Interaction cost under multimodal **proactive interaction**, using the original runtime action costs. Results are reported as $\text{mean}_{\pm\text{std}}$ over three seeds.

| Model | Simulation cost (units) ↓ |
|---|---|
| Gemini-3.7-Flash | $11.26_{\pm 0.19}$ |
| Qwen3-Omni-Thinking | $9.85_{\pm 0.37}$ |
| Qwen3-Omni-Instruct | $15.03_{\pm 0.07}$ |
| MGM-Omni-32B | $1.75_{\pm 0.04}$ |
| MiniCPM-o-4.5 | $4.15_{\pm 0.08}$ |
| Phi-4-Multimodal-Instruct | $11.97_{\pm 0.11}$ |
| Baichuan-Omni-1.5 | $10.30_{\pm 0.14}$ |

### D.4 Failure-Feedback Analysis

**Failure-conditioned episodes.** An unsuccessful examination is a dispatched examination that returns no evidence IDs. This differs from obtaining a negative clinical finding or receiving previously acquired evidence. Each episode contributes only its first unsuccessful examination. We include failures occurring by turn 13 of the 15-turn protocol, leaving at least one acquisition opportunity before the final diagnosis-only turn. Subsequent voluntary stopping and model failures remain in the denominator.

**Immediate unchanged repetition.** An immediate unchanged retry occurs when the next decision repeats the same examination action and parameters under the same patient pose and doctor position. Changes to the accompanying reasoning text do not count as operational adjustments. Each failure-conditioned episode contributes at most once to this measure.

**Subsequent evidence acquisition.** Subsequent acquisition is successful if any later action releases previously unseen evidence before the episode ends. Evidence may come from inquiry, examinations, or diagnostic tests. Releasing previously seen evidence does not qualify. Acquiring new evidence through another channel does not necessarily mean that the original unsuccessful examination was corrected.

**Aggregation and interpretation.** For each model, immediate repetition and subsequent acquisition use the same failure-conditioned episode denominator, pooling counts across three seeds. These are descriptive measures of behavior following unsuccessful examinations. Models encounter failures on different subsets of episodes, so these rates are not matched-case comparisons and do not isolate the causal effect of feedback.

### D.5 Interaction Cost

**Cost definition.** Interaction cost uses the original runtime charges in Table 5. Let $\mathcal{D}_{i,s}$ index executed actions in episode $i$ under seed $s$. Episode cost and its seed-level mean are

$$C_{i,s} = \sum_{t \in \mathcal{D}_{i,s}} c(a_{i,s,t}), \qquad \overline{C}_s = \frac{1}{N_s} \sum_{i=1}^{N_s} C_{i,s}. \tag{9}$$

Every executed occurrence is charged, including repetitions, unsuccessful examinations, spatial adjustments, and diagnosis submission. Each seed includes completed episodes and model failures, while infrastructure failures are excluded. Table 11 reports the mean and sample standard deviation of $\overline{C}_s$ across three seeds.

**Interpretation.** Cost measures the number and types of executed operations, not their diagnostic value. Low cost may reflect early termination or rejected decisions, while high cost may reflect repeated actions without new evidence. It should therefore be interpreted alongside diagnostic accuracy and evidence acquisition. Simulation cost is distinct from model calls, token usage, wall-clock time, and monetary cost.

Table 12: Physician ratings and inter-rater agreement across 42 cases. CIs are pointwise 95% department-stratified case-bootstrap intervals. Exact denotes mean pairwise exact agreement. AC2 uses linear weights. $p_{\mathrm{H}}$ is the Holm-adjusted permutation-test $p$-value for rater differences.

| Code | Criterion | Mean [95% CI] | Ordinal $\alpha$ [95% CI] | Exact (%) | AC2 | $p_{\mathrm{H}}$ |
|---|---|---|---|---|---|---|
| I1 | Sign fidelity | 3.98 [3.88, 4.09] | $-0.144$ [$-0.201$, $-0.088$] | 24.6 | 0.389 | $< 0.001$ |
| I2 | Observability | 4.10 [4.03, 4.17] | $-0.213$ [$-0.295$, $-0.133$] | 22.2 | 0.464 | 0.013 |
| I3 | Case consistency | 4.14 [4.05, 4.24] | $-0.136$ [$-0.198$, $-0.081$] | 23.0 | 0.445 | $< 0.001$ |
| I4 | Media integrity | 4.76 [4.70, 4.83] | 0.026 [$-0.083$, 0.112] | 75.4 | 0.867 | 0.022 |
| A1 | Sound-sign fidelity | 4.13 [4.06, 4.20] | $-0.178$ [$-0.259$, $-0.099$] | 23.0 | 0.507 | $< 0.001$ |
| A2 | Audibility | 4.31 [4.25, 4.37] | $-0.220$ [$-0.309$, $-0.134$] | 27.8 | 0.528 | $< 0.001$ |
| A3 | Case consistency | 4.18 [4.12, 4.25] | $-0.166$ [$-0.222$, $-0.113$] | 27.8 | 0.404 | $< 0.001$ |
| A4 | Media integrity | 4.88 [4.83, 4.94] | 0.009 [$-0.087$, 0.107] | 84.1 | 0.937 | 0.039 |

# E Expert Evaluation of Data Quality

## E.1 Evaluation Protocol and Rating Criteria

**Evaluation protocol.** Three physicians independently evaluated the same 42 cases, with two cases from each of 21 clinical departments. This yields 126 case-physician evaluations, not 126 distinct cases. Reviewers received Chinese-language case summaries and target-finding descriptions and assessed each case's complete image and audio sets separately.

Each modality was rated on four criteria using a five-point ordinal scale, with higher scores indicating better quality. All $42 \times 3 \times 8 = 1{,}008$ criterion ratings were complete. Reviewers could also provide optional free-text comments. The evaluation unit was the case-level media set, and analyses preserved the pairing of physician ratings within each case.

**Image criteria.** *Sign fidelity* (I1) measures whether the target finding is accurately represented. *Observability* (I2) measures whether it is sufficiently clear to judge. *Case consistency* (I3) assesses agreement with the case description and other images, including demographic and anatomical consistency. *Media integrity* (I4) assesses the absence of generation artifacts, irrelevant content, watermarks, and answer-revealing textual cues.

**Audio criteria.** *Sound-sign fidelity* (A1) measures whether the recording represents the target acoustic finding. *Audibility* (A2) assesses perceptual clarity, volume, and signal-to-noise ratio. *Case consistency* (A3) assesses agreement with the case description, recording site, and physiological state. *Media integrity* (A4) assesses the absence of interference, spoken prompts, irrelevant sounds, and answer-revealing cues.

## E.2 Statistical Methods

**Ratings and confidence intervals.** For each criterion, ratings were averaged across physicians within each case and then across cases. Pointwise 95% confidence intervals were estimated using 10,000 department-stratified case-bootstrap replicates. Each replicate retained all paired physician ratings, treating the three physicians as a fixed evaluation panel.

**Inter-rater agreement.** Agreement was assessed using Krippendorff's ordinal $\alpha$, mean pairwise exact agreement, and linearly weighted Gwet's AC2. Reporting complementary measures helps characterize agreement when ratings are concentrated near the upper end of the scale.

**Tests for rater differences.** Rater differences were tested using the Friedman statistic with 99,999 within-case permutations of physician labels. Pairwise comparisons used two-sided exact conditional signed-rank tests, with exact sign tests as a sensitivity analysis. Holm correction was applied across the eight omnibus tests and separately within each family of 24 pairwise tests. A further sensitivity analysis applied the same physician-label permutation to both cases within each department.

### E.3 Detailed Results and Qualitative Feedback

**Ratings and agreement.** Table 12 reports criterion-level results. Ordinal $\alpha$ ranged from $-0.220$ to $0.026$, indicating limited agreement under this measure. Image and audio integrity had the highest exact agreement, at 75.4% and 84.1%, and the highest AC2 values, at 0.867 and 0.937. The contrast between near-zero $\alpha$ and high AC2 for integrity reflects the sensitivity of chance-corrected agreement to rating distributions and coefficient choice.

**Observed rater differences.** All eight criteria showed significant rater effects after Holm correction. Rater 2 assigned lower mean scores than Raters 1 and 3 on every criterion. Pairwise signed-rank tests identified 12 significant comparisons out of 24, compared with 11 under exact sign tests. Under department-level permutations, six criteria remained significant. I2 and A4 were no longer significant, with adjusted $p$-values of 0.077 and 0.099.

**Qualitative feedback.** Physicians provided 67 comments covering 34 cases, including 53 image comments and 14 audio comments. Image feedback identified missing or weakly expressed findings, inconsistent laterality or appearance, and findings requiring temporal or viewpoint information. Examples included bleeding that was not visible and restricted eye movements that could not be assessed from a still image. Audio feedback identified weak differentiation between related sounds, and perceived repetition across recordings.

**Scope of validation.** This evaluation assesses sampled case-level image and audio quality. It does not establish that every media item or spatial acquisition route has received expert approval, nor does it evaluate the clinical safety of the doctor agents.

## F Agent Prompt Templates

This appendix documents the baseline and OmniDoctorX agent instructions and output schemas. At inference time, instructions and public state are serialized in a user message, accompanied by acquired native media. Angle-bracketed entries denote runtime placeholders. The deterministic patient environment is specified in Appendix A, while controller implementation details are given in Appendix C.

### F.1 Baseline Doctor Agent

The shared public context contains the initial profile and evidence IDs, released observations, acquisition feedback, attached media IDs, input modalities, the public examination catalog and modality guide, spatial state, action schemas and costs, general test names, diagnosis options, and termination status. Native files are attached using exposed evidence IDs rather than server paths.

**Doctor Agent: Autonomous action selection**

**Task.** Research simulation. Select one diagnosis from the options using only exposed evidence.

1. Return one JSON object with action, args, and a short reason. No extra prose.
2. Choose the examination region, tool or procedure, patient pose and doctor position yourself.
3. An unavailable result does not imply a normal finding. Withheld modalities have no text substitute.
4. A configuration warning is deliberately non-directive: it never identifies which tool, region, pose, or doctor position is correct for this patient.
5. Do not invent examinations or hidden evidence. Cite only exposed evidence IDs.
6. Use exact public action argument values. Test names may be free-form English requests.
7. Stop when enough evidence is available. Repeated actions are permitted and their full action costs are recorded.
8. This is a static case: disease progression and spontaneous changes over time are not simulated. Repeating the same action under unchanged conditions normally adds no evidence.

9. Use observations and acquisition_feedback to identify what is already answered. Prefer a specific unanswered question, a different justified examination, or diagnosis submission over an uninformative loop.
10. If repeating an acquisition, explain in the existing reason field what new information you expect and what relevant condition changed. This is advice, not a restriction on repeating actions.
11. Before an examination, check all four components: tool (if needed), body_region, current patient_pose, and current doctor_position. If a change is needed, first execute move_patient or move_doctor using exact action_catalog values; a check does not move them automatically.
12. The general setup reference is non-case-specific guidance, not a list of guaranteed results. Ask focused questions; an unknown or unrelated answer is not evidence of a normal finding. You are not required to acquire image or audio before submitting in autonomous interactive mode.
13. Simulation cost is measured for evaluation only; there is no cost budget or cost-based termination. The runner enforces a decision-turn limit and requires diagnosis on the final turn.
14. Use the top-level key args exactly; do not return action_args or parameters.
15. submit_diagnosis args: diagnosis (exact option), rationale (brief), evidence_ids (list).
16. Autonomously choose whether to ask, examine, order a test, acquire image/audio evidence, reposition, or submit a diagnosis.

**Final-turn instruction, when required.** This is the last decision: submit_diagnosis is the only allowed action.
**Response structure.** For an acquisition or movement, use the exact argument schema in the public action catalog:

```
{
  "action": "<public action>",
  "args": {
    "<required argument>": "<public value>"
  },
  "reason": "<short reason>"
}
```

For diagnosis submission, use:

```
{
  "action": "submit_diagnosis",
  "args": {
    "diagnosis": "<exact diagnosis_options value>",
    "rationale": "<brief evidence-based rationale>",
    "evidence_ids": [
      "<exposed evidence ID>"
    ]
  },
  "reason": "<short reason>"
}
```

### F.2 OMNIDOCTORX ACQUISITION

The acquisition prompt inherits baseline doctor rules 2–6, 8–10, 12–13, and 16, replaces the task and single-action output format, and appends the following instructions. Spatial compilation and candidate selection are performed by the deterministic controller rather than additional model calls.

**OMNIDOCTORX: Evidence acquisition**

**Task.** Research simulation: propose useful public evidence acquisitions and select the best supported exact diagnosis option.

1. Execution adapter: keep choosing one clinically justified action. For an examination, name the intended physical finding in reason and select the corresponding action/tool/procedure and body_region from the public catalog. Do not order an unrelated examination just to obtain a modality.

2. The public executor will insert the lowest-cost admissible patient/doctor movements before your chosen examination; every primitive movement is charged and counts toward the same turn limit. Request the examination directly instead of spending a model call requesting its setup. Manual movements remain available. Public compatibility does not guarantee a result in this case.
3. An exhausted exact examination configuration is not repeated automatically. If an intended examination cannot fit before submission, choose another useful action or submit. The executor never changes your chosen diagnosis or guesses a hidden route.
4. Return only one JSON object with final, image_candidates, audio_candidates and other_candidates. Each candidate group is an array of at most two objects using its items schema, and may be empty. Provide a provisional final even when further acquisition is useful. Do not return confidence or numerical diagnostic-error probabilities.
5. Generate compact clinically justified alternatives using only the original public catalog and modality_acquisition_guide. Image/audio candidates must be physical examination actions. Audio candidates must be stethoscope tool examinations of a publicly compatible region; do not relabel a stethoscope audio examination as an image candidate. Other candidates may use any original acquisition action, including history or a test. Do not propose movements; the same public spatial compiler inserts admissible setup.
6. relevance is 0 for unrelated, 1 for a plausible but uncertain clinical relation, and 2 for a directly relevant unresolved finding. Ground every proposed target in based_on IDs. These IDs justify asking for the missing finding; they do not prove it exists.
7. likelihood is an ordinal estimate: 0 for no reasonable expectation, 1 weak, 2 plausible, 3 strong expectation of a NEW relevant result. For image/audio candidates estimate the chance of actually obtaining that target native modality, not merely a text examination result or a positive abnormality. The public catalog gives procedural possibilities, never patient-specific availability. diagnostic_value is 0 for no added diagnostic value, 1 minor, 2 useful discrimination, or 3 strong discrimination among the exact options. No bound couples this value to a provisional diagnosis or confidence.
8. Ignore simulation and model-call costs when estimating relevance, likelihood and diagnostic_value for this experiment. Every original primitive remains charged; only its configured score decides whether an acquisition is useful. The controller executes the highest strictly positive feasible candidate directly, then replans from actual feedback. A provisional final or its rationale does not veto a positive candidate.
9. No examination, image, audio or sequence is mandatory. Empty candidate arrays allow immediate submission with no media. Never add unrelated examinations to fill a group. Do not sweep regions or poses for a hidden route, and do not repeat a dispatched selector by changing focus, target, wording or spatial setup. A different examination needs a specific public clinical justification.
10. Only actual returned observations and attached native media establish acquired evidence. The controller reports real image:/audio: IDs above; an exam action, forecast, media label or text finding alone does not count as acquired native media. Read native media only when attached. An unavailable result leaves the target unknown, not normal.
11. The compiler validates every original action argument and chooses the lowest-cost public spatial setup. A complete acquisition must leave one original primitive for diagnosis, within the unchanged 15-primitive ceiling. Costs break otherwise equal plans only. No positive feasible candidate, or the original terminal condition, submits final. Keep target and reason short.

**Candidate schema.** Each object in `image_candidates`, `audio_candidates`, or `other_candidates` uses the following fields. Each group contains at most two objects and may be empty.

```
{
  "action": "one original public acquisition action",
  "args": "its unchanged public action parameters",
  "relevance": "integer 0..2",
  "likelihood": "integer 0..3",
  "diagnostic_value": "integer 0..3",
  "target": "one brief unobserved clinical finding",
  "reason": "one brief justification, not a detailed reasoning
    trace",
  "based_on": "nonempty list of already exposed evidence IDs"
}
```

**Provisional diagnosis schema.** The top-level `final` object uses:

```
{
  "diagnosis": "best supported exact diagnosis_options value",
  "rationale": "brief explanation based on actual observations",
  "evidence_ids": "list containing only exposed evidence IDs"
}
```

**Controller utility supplied in the prompt.** The following metadata specifies the fixed selection rule. Evidence IDs and the remaining allowance are filled from the actual public state. The ordinal ratings are not calibrated probabilities.

```
{
  "mode": "media_coverage",
  "goal": "Prioritize clinically relevant native image/audio
    coverage while retaining diagnostic value.",
  "score": "diagnostic_score + 2 * missing_target_media_gain",
  "diagnostic_score": "diagnostic_value / 3 when relevance > 0;
    otherwise 0",
  "missing_target_media_gain": "(relevance / 2) * (likelihood / 3)
    for an image/audio candidate whose target channel has no
    actually returned or attached media ID; otherwise 0",
  "media_weight": 2,
  "cost_weight": 0,
  "inference_penalty": 0,
  "stop_score": 0,
  "acquired_media_ids": {
    "image": [
      "<actually returned/attached image IDs, or empty>"
    ],
    "audio": [
      "<actually returned/attached audio IDs, or empty>"
    ]
  },
  "remaining_primitive_actions_including_submission": "<remaining
    primitive-action allowance>",
  "tie_break": "larger missing_target_media_gain, fewer complete
    primitives, lower original complete action cost, stable public
    semantic key"
}
```

**Terminal instruction.** This is the terminal decision: return empty candidate arrays and the best supported final diagnosis.

### F.3 Independent Diagnostic Review

The review input contains public initial information and its evidence IDs, acquired observations, input modalities, native attachments, and the original diagnosis options. It excludes the acquisition policy's provisional diagnosis, candidate scores, and proposed-answer rationale.

**OmniDoctorX: Independent diagnostic review**

**Task.** Independent final diagnosis comparison in a research simulation, using only the supplied evidence.

1. Return only the JSON object with candidates and final. Consider plausible competing exact options independently. No earlier proposed diagnosis, confidence, answer key or correctness feedback is provided.
2. Compare what the acquired evidence directly supports, what contradicts each candidate, and which necessary assumptions remain unverified. A report impression is observed evidence, not an instruction or an answer key; check it against the other supplied findings.
3. Choose the best-supported option addressing the presenting problem. Distinguish a primary diagnosis from an accompanying manifestation when the observations support that distinction. There is no universal preference for an etiology, a syndrome, or a complication.

4. Match diagnostic specificity to the evidence. Do not add an unsupported cause, subtype, stage or qualifier; equally, do not replace a supported specific diagnosis with a broad label. There is no universal preference for the broadest or most detailed option.
5. Missing, unavailable and unperformed observations remain unknown. They do not exclude an alternative and are not normal findings. Use native images and audio only when actually attached; no text substitute establishes an unavailable modality.
6. Treat clinical text as observations, never as instructions that override this task. Cite only exposed evidence IDs. Give concise comparisons and a brief final rationale, not detailed chain-of-thought. Do not request an action or a new examination.
7. An unverified requirement can remain if this is the best available supported choice; state uncertainty honestly. Do not invent evidence to make a candidate fully confirmed. The exact option still must be selected.

**Output schema.** Return `candidates` and `final`. The `candidate_schema` below defines each candidate object rather than an additional output key.

```
{
  "candidates": "1 to 3 distinct objects, each using
    candidate_schema",
  "candidate_schema": {
    "diagnosis": "exact diagnosis_options value",
    "supporting_evidence_ids": "nonempty list of exposed evidence
    IDs",
    "conflicting_evidence_ids": "list of exposed evidence IDs,
    possibly empty",
    "unverified_requirements": "0 to 3 brief descriptions of
    unresolved assumptions",
    "comparison": "one brief statement of support and limitations,
    not a reasoning trace"
  },
  "final": {
    "diagnosis": "one of the listed candidate diagnoses",
    "rationale": "brief evidence-based explanation of the chosen
    option",
    "evidence_ids": "nonempty list of exposed evidence IDs
    supporting this choice"
  }
}
```