MEDIMAGEOSWORLD: BENCHMARKING GUI AGENTS FOR MEDICAL IMAGE CONSOLES

Anonymous authors
Paper under double-blind review

ABSTRACT

Graphical consoles offer a practical interface for medical acquisition assistance, allowing agents to work through the controls and visual feedback used by human operators. Reliable assistance requires linking on-screen anatomy to acquisition decisions that determine what image evidence becomes available next. We introduce MedImageOSWorld, a benchmark for evaluating this capability in simulated CT, MR, and ultrasound consoles. Using screenshots and mouse-and-keyboard actions, agents configure protocols, plan acquisitions, inspect the resulting images, and make corrective adjustments across seven capability levels, from console operation to feedback-driven control. Evaluation combines task-specific workflow checks with hidden anatomical ground truth to assess procedural completion and acquisition outcomes separately. A common evaluation protocol specifies episode conditions and interaction budgets, while recorded trajectories support analysis of how agents observe, act, and respond to acquisition feedback. Across eleven open-weight agents, success rates range from 3.0 to 25.0 on a 0–100 scale while workflow-progress rates reach 21.5–74.5: agents complete much of the console workflow but rarely acquire the intended anatomy. Success collapses between perception and planning, from 79–90% at the lowest three levels for the best agent to at most 12% for millimetre-level planning and 0% for closed-loop control. Two proprietary agents reach success rates of about 34 and exceed the best open-weight agent mainly in acquisition quality (61 versus 42). MedImageOSWorld provides a controlled setting for studying whether general-purpose GUI agents can translate visual observations into effective medical acquisition decisions.

1 INTRODUCTION

Medical image acquisition depends on operators translating an examination request into console actions. They configure protocols, plan anatomical coverage or imaging planes, and assess the resulting images. The console is the interface where an operator turns an examination goal into acquired image evidence. An inadequate acquisition may require adjustment, repetition, or escalation.

The graphical console is a natural interface for acquisition assistance: it brings protocol settings, spatial planning, and image assessment into the operator's existing workflow. Agents that use screenshots and standard input actions can interact through this interface without requiring a dedicated scanner-control API or direct access to internal acquisition state. This offers a route to automation across heterogeneous console software while keeping actions and visible outcomes available for operator review. It also creates a demanding control problem: the agent must connect what is shown on screen to acquisition decisions whose consequences appear in subsequent images.

Recent advances in general-purpose GUI agents make this interaction setting increasingly relevant. Computer-use benchmarks evaluate multimodal agents on stateful workflows over desktop applications and web services (Yuan et al., 2026). Medical acquisition poses a further question: can such agents combine visual understanding, spatial reasoning, and feedback-driven control to acquire the intended anatomy without human correction during an episode? We ask not only whether an agent can operate the machine, but whether its actions produce an acquisition that meets the intended goal.

The distinction arises because acquisition actions determine what evidence becomes available for subsequent decisions. We call this property *action-conditioned acquisition*: an agent action alters the acquisition process and thereby changes the medical-image evidence produced at subsequent

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.

steps. The agent must identify the anatomy on screen, relate image coordinates to acquisition geometry, and use the resulting evidence to decide whether to proceed or adjust. Changes to geometry, timing, or probe pose affect acquisition; display-only adjustments do not. An agent can complete the prescribed workflow and still acquire the wrong anatomy. Workflow completion alone does not establish that the intended anatomy was acquired.

Existing environments address parts of this evaluation problem. Clinical benchmarks cover electronic health records, simulated encounters, and administrative workflows (Jiang et al., 2025; Schmidgall et al., 2026; Bedi et al., 2026), while MedCUA-Bench evaluates clinical GUI interaction with safety-critical errors (Yu et al., 2026). These settings primarily concern records, software state, or predefined observations rather than imaging acquisition. ABRA, RadABench, and MedSPOT study post-acquisition viewing, radiology tool use, and interface grounding (Maksudov et al., 2026; Zheng et al., 2024; Shakeel et al., 2026); they do not let the agent control the acquisition that generates the underlying study. Autonomous MRI and MaRGE/MaRCoS address acquisition and scanner control (Ravi & Geethanath, 2020; Algarín et al., 2024), but are MRI-specific systems rather than standardized benchmarks for general-purpose GUI agents. The remaining need is an evaluation setting that connects executable console workflows to action-conditioned acquisition and independent anatomical ground truth.

To study this question, an environment must preserve the dependencies between planning, acquisition, and assessment while retaining ground truth for the acquired result. We introduce **MedImageOSWorld**, a benchmark with three executable consoles: computed tomography (CT), magnetic resonance (MR), and ultrasound. All three derive observations from XCAT-based digital human anatomy phantoms (Segars et al., 2010). The environment couples an order, a screening record, persistent patient and device state, a protocol library, and an action-conditioned acquisition backend. Agents select and configure protocols, plan anatomy-dependent slices, volumes, or views, acquire images, and assess their adequacy. They can re-plan, re-acquire, repeat a view, or escalate. Figure 1 separates the three stages that define an episode: visible console interaction, a hidden physical-state transition, and outcome evaluation against anatomical ground truth.

MedImageOSWorld provides a setting for studying how general-purpose agents combine visual understanding, spatial reasoning, and feedback-driven control in tasks with simulated physical consequences. A shared capability ladder organizes these requirements and helps locate failure points. Evaluation separates workflow completion from acquisition quality using hidden ground truth in the digital subject's coordinate frame. This makes the target of evaluation the connection between interface actions and measurable acquisition outcomes, rather than procedural completion alone. Our contributions are:

- **Action-conditioned acquisition as an evaluation setting.** We formalize medical image acquisition as a GUI-agent task in which actions change the physical acquisition state and therefore the image evidence observed next. This setting separates two questions that existing benchmarks conflate: whether the agent completed the console workflow, and whether the resulting acquisition captured the intended anatomy.
- **MedImageOSWorld, a benchmark of executable scanner consoles.** Three simulated consoles (CT, MR, ultrasound) expose protocol selection, geometric planning, acquisition, review, and corrective actions through rendered screenshots and a constrained mouse-and-keyboard interface, with no access to interface or simulator state.
- **Ground-truth scoring in the subject's coordinate frame.** Each committed acquisition is linked to the digital subject's coordinate frame, so a hidden oracle can grade geometry and anatomical coverage in millimetres, degrees, and coverage fractions without human assessment. Success, workflow progress, and acquisition quality are reported as three separate 0–100 scores, with environment faults kept apart as `INVALID`.
- **An empirical study of current GUI agents.** We evaluate 11 open-weight models and two API-served proprietary models on 31,197 valid episodes included in scoring (30,513 for open-weight agents and 684 for proprietary agents) under a frozen interaction and scoring protocol. The best open-weight agent, Qwen3.8-27B, reaches a success rate of 25.0 but a workflow-progress rate of 74.5; no open-weight agent exceeds 12% success on millimetre-level planning, and CT remains nearly unsolved, with success at or below 1.5 for every open-weight agent. The simulator and data will be available on Anonymized Hugging Face.

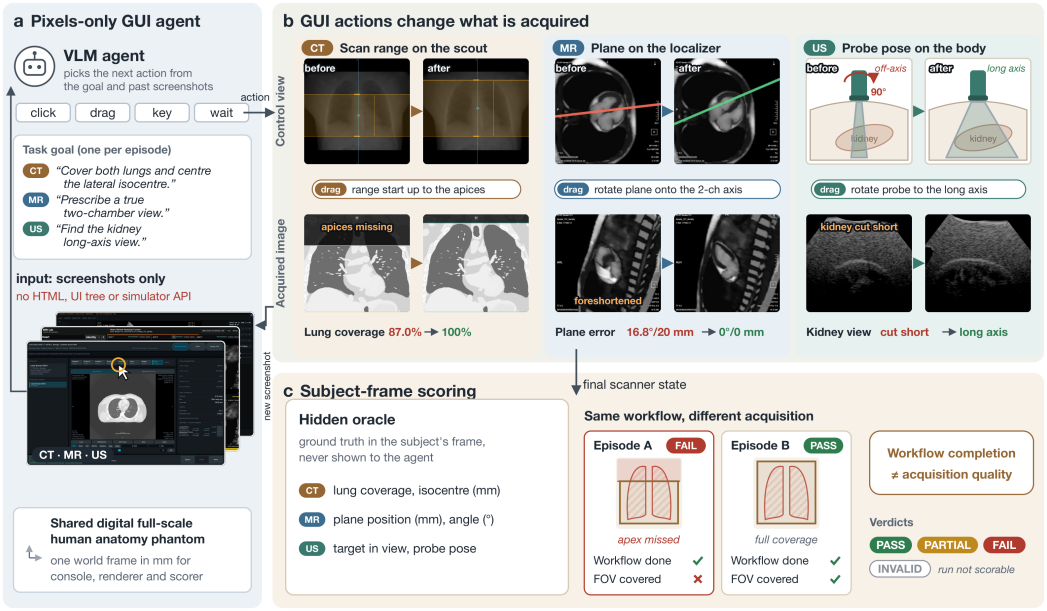


Figure 1: **MedImageOSWorld evaluation loop.** (a) A GUI agent receives console screenshots and acts through mouse and keyboard inputs, without interface or simulator state. (b) Changes to CT scan range, MR plane, or ultrasound probe pose alter the image evidence acquired next. (c) A hidden oracle evaluates the final acquisition in the subject's coordinate frame, separating workflow completion from acquisition quality. The displayed measurements illustrate the scoring interface and are not experimental results.

2 Related Work

General computer-use benchmarks. Table 1 contrasts representative clinical, radiology, and acquisition environments across the interaction, acquisition, and evaluation properties that define this task setting. OSWorld evaluates multimodal agents by executed outcomes in real computer environments, and OSWorld 2.0 extends this line to longer-horizon tasks (Xie et al., 2024; Yuan et al., 2026). AndroidWorld varies task parameters in a dynamic mobile environment, while WebArena evaluates functional correctness on realistic web tasks (Rawles et al., 2025; Zhou et al., 2024). Like these benchmarks, we score executed outcomes rather than action traces and use task-specific parameter variation; MedImageOSWorld additionally introduces a physical acquisition state that determines the next observation. Validation coverage is reported for the checks actually performed rather than assumed for every planned control or perturbation.

Clinical and radiology agents. MedAgentBench and AgentClinic study EHR use and simulated encounters (Jiang et al., 2025; Schmidgall et al., 2026); the health administration benchmark and MedCUA-Bench cover administration and clinical GUI interaction (Bedi et al., 2026; Yu et al., 2026). RadABench, MedSPOT, and ABRA study radiology tool use, interface grounding, and DICOM viewing (Zheng et al., 2024; Shakeel et al., 2026; Maksudov et al., 2026). In these settings, the relevant records or imaging study are fixed before the episode; the agent does not control how the evidence is acquired. MedImageOSWorld adds this acquisition loop before review or interpretation.

Autonomous acquisition and digital subjects. Autonomous MRI supports automated MRI protocolling, planning, acquisition, and reconstruction, while MaRGE/MaRCoS supports research MRI control (Ravi & Geethanath, 2020; Algarín et al., 2024). These are MRI-specific acquisition systems rather than benchmarks for general-purpose GUI agents. MedImageOSWorld links console actions and committed acquisitions to the subject's coordinate frame so that geometry and anatomical coverage can be scored automatically. The modality-specific scope and omissions of the renderers are documented separately rather than treated as a general claim of scanner-level physical fidelity.

Table 1: Comparison of environments related to GUI agents and medical image acquisition.

Work	Setting	Pixels-only obs.	Plans acq.	Action-dep. image	Subject-frame GT
OSWorld 2.0 (Yuan et al., 2026)	Desktop and web computer use	✓	—	—	—
MedAgentBench (Jiang et al., 2025)	FHIR-based EHR tasks	—	—	—	—
RadABench (Zheng et al., 2024)	Prompt-driven radiology tools	—	—	—	—
ABRA (Maksudov et al., 2026)	DICOM viewer and reporting tools	○	—	—	—
MaRGE/MaRCoS (Algarín et al., 2024)	Research MRI scanner console	—	✓	✓	—
Autonomous MRI (Ravi & Geethanath, 2020)	Automated brain MRI workflow	—	✓	✓	—
MedImageOSWorld (ours)	**Simulated CT, MR, and US consoles**	✓	✓	✓	✓

Notes. "Action-dep. image" means that the agent's choices affect subsequently acquired image evidence, beyond changing the GUI display or navigating a fixed study. ✓: full support; ○: partial or indirect support; —: absent or not applicable. ABRA provides image pixels but controls the viewer through function-calling tools, so its observation is not pixels-only.

3 THE MEDIMAGEOSWORLD ENVIRONMENT

We model each episode as a partially observed interaction between an agent and a simulated console. The console state has two components. The *physical acquisition state* determines what the scanner would acquire, such as the CT slab and isocentre, the MR imaging plane, or the ultrasound probe pose and tissue contact. The *interface and workflow state* records the worklist, sequence queue, workflow stage, and display settings. At each step, the agent issues a single GUI action; the simulator then updates the console state using task-relevant physical abstractions rather than a full scanner model. Display-only actions, such as adjusting window width and level, modify the interface state but leave the physical acquisition state unchanged. Section 3.3 describes the separate evaluation of workflow completion and acquisition quality. Figure 1 illustrates this loop.

3.1 OBSERVATION, ACTION, AND TIME

The primary *observation* is a PNG screenshot of the console rendered at a fixed viewport. The *action* space consists of constrained mouse events and an allowlisted set of keyboard events, so the agent interacts through the same visible console and standard input modalities available to a human operator. The agent receives no DOM, accessibility tree, simulator or interface state, script-execution primitive, hidden UI-quality text, or direct scanner-control API; task geometry and the consequences of each action are therefore available only through the rendered pixels. Before execution, a declared, version-fixed, deterministic codec may normalize model output into a legal action. For every step, an action receipt records the raw model output, any normalization or repair applied, whether the action was executed or rejected, and any additional feedback returned to the agent. Rejected actions consume steps as specified by the frozen interaction profile, and any model-specific repair logic or additional prompting is disclosed as part of the agent configuration. Simulated *time* advances only through executed actions, including explicit wait actions, rather than during model inference.

3.2 THREE CONSOLES

All three consoles share a common structure (Figure 1b): planning controls modify the physical acquisition state, acquisition commits it, and the acquired images provide evidence for assessment and correction. **Computed tomography (CT).** After worklist and protocol selection, the agent plans on a scout image by setting the scan slab, isocentre, monitoring region of interest (ROI), and acquisition timing. The acquired stack is presented separately from the prescription, so undercoverage

can be detected and corrected. Task families cover thoracic, renal, hepatic, and cardiac acquisition, including contrast timing, repeat acquisition of clipped anatomy, and coverage under a dose-index constraint. **Magnetic resonance (MR).** The agent prescribes imaging planes and volumes on localizer images by setting plane position, orientation, field of view, and slab or slice coverage; each acquisition reslices the digital subject at the committed geometry. Within the sequence queue, acquired images guide the refinement or repetition of subsequent prescriptions for cardiac cine planes, cardiac T1/T2 mapping, and bilateral renal coverage. **Ultrasound (US).** The live renal image is rendered from the current probe position, orientation, and tissue contact, and from the depth, frequency, and focus settings. Tasks require searching for a kidney view, adjusting acquisition parameters, distinguishing contact loss from inadequate gain, and freezing and saving a selected frame.

3.3 Digital subjects and ground truth

The digital anatomy phantoms provide registered anatomical structures in a shared subject coordinate frame, in which every planning control and committed acquisition is expressed. CT slabs and isocentres, MR planes and volumes, and ultrasound probe planes can therefore be compared directly with hidden anatomical masks and landmarks, enabling automatic measurement of endpoint and centre error, plane position and orientation, field of view, and anatomical coverage, independent of screen coordinates or click sequences.

Workflow checks use the interface state and event log to verify that required stages and actions occurred, whereas the hidden oracle evaluates the anatomy actually acquired; a workflow can thus complete without meeting the acquisition-quality criteria. Rendering assumptions, coordinate conventions, and modality-specific scope are documented in the supplementary section *Simulator Rendering and Physical-Model Scope*.

4 Tasks and Capability Ladder

Task sourcing. We first defined the capability ladder and then designed tasks against it (Figure 2a). Three medical imaging experts, assisted by a frontier AI model, proposed, cross-reviewed, and finalized each task. The benchmark contains 49 task definitions: 8 for CT, 34 for MR, and 7 for ultrasound. They cover CT coverage planning, dose constraints, and undercoverage recovery; MR console operation, acquisition planning, mapping, and coverage assessment; and ultrasound kidney operation, measurement, search, optimization, and fault recovery. The tasks evaluate acquisition operations, not diagnostic accuracy. The supplementary section "Task Catalog and Reproducibility Details" lists every task with a short description and its capability level; execution coverage is reported separately from task membership.

Capability ladder. A task's level is the capability it targets, not the interaction mechanism it uses (Figure 2a). L0–L1 concern operating the console, L2 concerns interpreting the current observation, and L3–L6 add spatial quantification, physical constraints, temporal closure, and recovery. A composite task is counted once, at its primary level; for example, `ct_ui_smoke_routine` completes the full routine workflow but is placed at L3 because it is gated on lung geometry. The ladder helps locate where performance breaks down, but it is not a psychometrically validated scale, and a single failure does not identify a unique cause: a planning failure may originate in an earlier perception error. Registered coverage differs by modality; a zero means that the level is explicitly empty, not unknown. Legacy level-like prefixes in task IDs do not determine the paper capability level; the registry and task catalog do.

Task specification. Each task is defined by its goal text, modality, capability level, reset distribution, observation and action profile, step and wall-time budgets, termination conditions, ground-truth scorer, and task and scorer versions. A task may define multiple conditions, and each condition targets 30 digital human subjects with three seeds per subject. Proprietary agents are evaluated on the first five assets of each task in the frozen order. The episode manifest distinguishes task definitions, conditions, planned slots, attempts, and scored episodes, and records the subject, seed, bundle, and reset information associated with each.

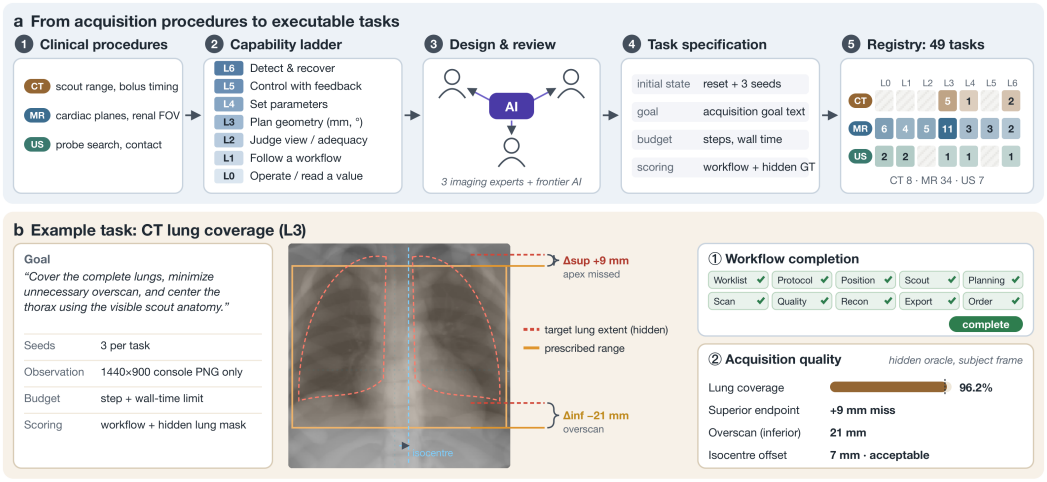


Figure 2: **From acquisition procedures to executable benchmark tasks.** (a) The shared L0–L6 capability ladder is defined first; three medical imaging experts, assisted by a frontier AI model, propose, cross-review, and finalize each task. Each task specifies an initial state, an acquisition goal, an interaction budget, and scoring conditions. (b) A CT lung-coverage task links scout-based planning to the acquired volume, with workflow completion and anatomical coverage scored separately. Numbers in (b) are illustrative.

5 Evaluation Protocol

Verdicts. The task scorer assigns `PASS`, `PARTIAL`, `FAIL`, or an undetermined outcome. Environment, receipt, or scorer faults are `INVALID`. Slots left unrun after a confirmed model-side early stop are counted as failures in SR. `INVALID` episodes and slots lost to infrastructure faults are excluded from metric denominators and reported separately.

Workflow and acquisition are scored separately. Workflow checks cover required and forbidden events and the terminal state. Acquisition quality is measured against hidden ground truth in the digital subject's frame: geometry in millimetres and degrees, and coverage as the fraction of the target structure inside the acquired volume. Coverage is computed from the acquired slab, never inferred from the prescribed range or a completion button (Figure 2b). Thresholds are task- and scorer-version specific; for the MR `L4_2ch` task, the current relaxed engineering profile gates centre, normal, field of view, and landmark errors at 8 mm, 5 degrees, 10 mm, and 5 mm. Continuous components remain visible alongside every verdict.

Protocol freeze and aggregation. A valid comparison freezes the task version, subject or bundle, reset, observation and action profile, budgets, scorer and thresholds, runtime, and agent configuration. Task-side manifests are shared across compared agents; agent-side prompts, codecs, deterministic normalizations, repair policies, and extra feedback may differ and are disclosed. Step accounting follows the frozen profile for rejected outputs, repairs, and corrective feedback; transport-level retries are logged separately. Results are laid out as model $\times$ modality $\times$ level $\times$ observation condition, with empty modality–level cells shown as N/A. We report three 0–100 metrics. Success rate (SR) is the percentage of registered slots that satisfy the task's complete success criterion. Progress rate (PR) is the mean percentage of registered workflow progress completed per slot. Acquisition quality (AQ) is the mean registered, normalized quality of the acquired result, computed from task-specific hidden-ground-truth measures. The legacy `PARTIAL` verdict provides no fractional credit to any main metric. Task-specific geometry, coverage, gate values, exhausted-budget rates, and the counts and reasons for `INVALID` episodes are reported alongside SR, PR, and AQ.

For tasks with multiple conditions, condition-specific results are reported. The MR renal-adequacy task's clean and defect conditions each target 30 subjects $\times$ 3 seeds, or 180 planned slots per model; this is not a claim of 180 valid scored episodes. Both conditions, answer frequencies, no-answer

rates, and actual denominators are reported. Conditions are weighted equally within each task, so the clean and defect conditions of MR renal adequacy each contribute half of its task score. Intervals resample subjects with their repeated runs grouped. The official scorer identity is the tuple of scorer code digest, oracle digest, and threshold version; changing any element creates a new scorer version.

Overall score. Let s_{ij} be a model's SR proportion for task j at level i, after condition aggregation and frozen denominator rules. The L0–L6 aggregate uses $w_i = i + 1$:

$$\text{Score} = 100 \frac{\sum_{i=0}^{6} (i+1) \sum_{j=1}^{n_i} s_{ij}}{\sum_{i=0}^{6} (i+1) n_i}, \qquad 0 \le s_{ij} \le 1. \tag{1}$$

Here n_i counts task definitions; the current denominator is 181. Higher levels receive greater preselected weight, not empirically validated psychometric difficulty. Modality and level scores are also reported. The pooled SR score is not modality-equal. SR, PR, and AQ are reported side by side and never merged.

6 Experiments

6.1 Setup

We ask four questions. Which registered requirements can current GUI agents satisfy across modalities and capability levels? How often does workflow completion coincide with acquisition quality? At which observation or corrective gate do non-passing episodes stop making progress? How sensitive are scores to the agent harness when the task-side protocol is held fixed?

We evaluate eleven open-weight GUI and vision-language agents: Qwen3.8-27B, UI-Mate-27B, GUI-Owl-1.5-32B-Instruct, GUI-Owl-1.5-32B-Thinking, Qwen3-VL-30B-A3B-Instruct, Qwen3-VL-30B-A3B-Thinking, UI-TARS-1.5-7B, Qwen2.5-VL-72B-Instruct, OpenCUA-32B, Gemma-4-31B, and Muse-Glimmer-30B. All are served locally with greedy decoding (temperature 0). We additionally evaluate OpenAI GPT-6 Sol (`gpt-6-sol`) and Anthropic Claude Opus 5.5 (`claude-opus-5-5`) through their provider APIs.

The target design assigns each task condition 30 subjects $\times$ 3 seeds. The MR renal-adequacy task has two conditions and therefore 180 planned slots per model, so 49×90 is not a unique suite total. The frozen manifest reports actual attempts and denominators, including early stops, cancellations, `INVALID`, undetermined, unscored, and unrun slots. Agent-side prompts, codecs, repairs, and feedback are recorded by condition.

Model and runtime revisions, hardware allocation, prompts, action codecs, normalization and repair policies, feedback, sampling, timeouts, and retries are reported in the supplementary configuration table under "Prompts, action codecs, and reference access."

6.2 Main results

Figure 3 summarizes performance across tasks and capability levels. Panel (a) reports each model's overall SR, PR, and AQ, with SR as the primary metric and PR and AQ shown as separate auxiliary measures. Panel (b) reports mean task-level SR within each capability level, pooled across CT, MR, and ultrasound. Overall SR uses Equation 1 across 49 tasks with a weighted denominator of 181.

Among open-weight agents, Qwen3.8-27B leads with an SR of 25.0, followed by Muse-Glimmer-30B (20.8), Gemma-4-31B (14.6), and UI-Mate-27B (10.7); the remaining seven score between 3.0 and 7.7. Difficulty follows the ladder: the best open-weight agent reaches 79–90% at L0–L2 but at most 12% at L3, and every open-weight agent scores 0% at L5. The proprietary agents reach SRs of 34.5 (GPT-6 Sol) and 34.1 (Claude Opus 5.5); their workflow progress is only 9–11 points above Qwen3.8-27B (85.5 and 83.4 versus 74.5), but their acquisition quality is about 19 points higher (61.1 and 61.3 versus 42.0), and CT remains difficult even for them (SR 10.3 and 1.7).

Workflow progress is high for the strongest agents, but acquisition quality lags well behind it. The API-served agents exceed Qwen3.8-27B in PR by only 9–11 points (GPT-6 Sol 85.5, Claude Opus 5.5 83.4, Qwen3.8-27B 74.5). Their lead in AQ is about 19 points (61.1, 61.3, and 42.0).

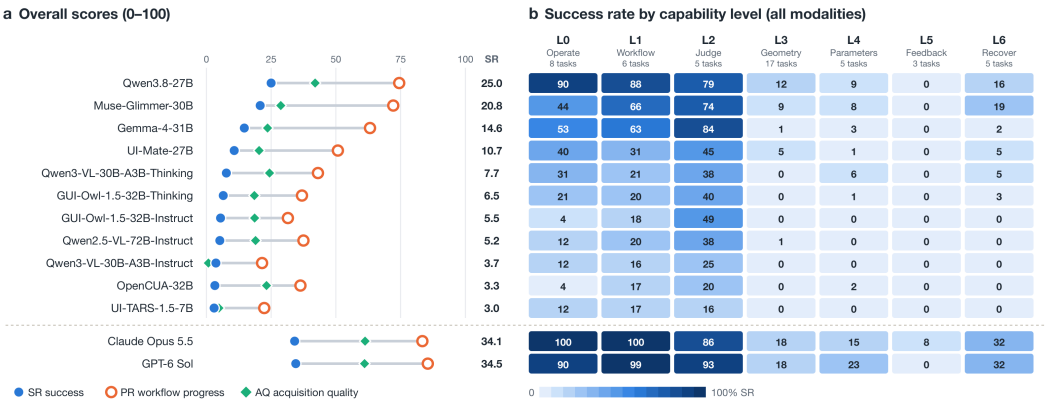


Figure 3: **Performance across tasks and capability levels.** (a) Overall success rate (SR; filled blue circles) is the primary metric: each scored episode receives 1 for `PASS` and 0 otherwise. Workflow progress (PR; open orange circles) measures completed workflow stages on a five-point scale (0, 0.25, 0.5, 0.75, or 1), without assessing acquisition quality. Acquisition quality (AQ; green diamonds) is a continuous score from 0 to 1 for the 28 eligible geometry and parameter tasks. All three measures are shown as percentages; PR and AQ are auxiliary measures and are not combined with SR. Overall scores use capability-level weights $w_i = i + 1$ across 49 tasks (weighted denominator 181). (b) Mean task-level SR within each capability level, pooled across CT, MR, and ultrasound; column headings show task counts. The divider separates locally served open-weight models from API-served proprietary models. API results use the first five assets per task.

The CT console illustrates the gap most directly. Their acquired volumes had median target coverage of 0.96 and 0.78 and median lateral centring errors of 6.7 mm and 14.3 mm: close to the planning thresholds, but outside them. Scaling the task tolerances from $0.5\times$ to $2\times$ changes the CT pass rate. In the CT recovery tasks, every agent accepted the repeat prompt and closed the study without correcting the clipped range (0 causal range corrections in the 360 Qwen3.8-27B and UI-Mate-27B episodes).

Coverage alone is also insufficient. On renal MR L6, Claude Opus 5.5 reached full bilateral coverage in all 12 translation-defect cells but passed none. In each case it restored coverage through a mechanism other than the responsible setting, so the causal-change gate failed.

6.3 Where agents fail

We decompose each scored trajectory into observation gates and corrective gates. A review gate records required viewing actions, such as opening an image and visiting specified slices or windows; it does not score interpretation. In renal tasks, `causal_change` records a change to the responsible acquisition parameter, and an independent coverage gate measures the resulting anatomy. Other corrective gates cover geometry, parameters, and terminal state.

Agents observe but do not correct. Observation gates are almost never the blocker. In renal MR tasks, agents review the initial series or scouts in 100% of episodes, and Claude Opus 5.5 satisfies every review gate in all 150 renal cells. Corrective gates are rarely met. Across the eleven open-weight models (about 990 valid episodes per renal task), the results are as follows:

- In numeric coverage correction, agents change the responsible parameter in 14% of episodes, and only 57% of those changes restore bilateral coverage.
- In coverage recovery, the corresponding rates are 7% and 36%.
- Replacement acquisitions are completed in 25–50% of episodes, but bilateral coverage is achieved in only 6–32%.

On the CT console, 1,032 of 3,084 valid episodes (33%) never committed an acquisition. Among early-stopped arms, the share of episodes reaching a committed acquisition ranged from 15%

(OpenCUA-32B) to 49% (Muse-Glimmer-30B). Gemma-4-31B and Muse-Glimmer-30B exported a study in 42% and 47% of episodes and still passed none.

Budget exhaustion and looping. Overall, 64.6% of scored open-weight episodes end with an exhausted action budget; these account for 75.7% of non-`PASS` episodes. By modality, the share of episodes ending this way is 70.9% in MR, 38.2% in CT, and 29.0% in ultrasound. On renal MR, exhaustion rises from 50% at L2 to 84%, 80%, and 75% at L3, L4, and L6. On CT, four agents repeated one identical click for at least 200 of 400 steps in most episodes: Qwen2.5-VL-72B (168/168), UI-TARS-1.5-7B (249/264), Qwen3-VL-30B-A3B-Instruct (193/216), and OpenCUA-32B (125/228). These four agents exported in at most 2% of episodes. On ultrasound, the three zero-pass agents exhausted the 64-event budget in 89–100% of episodes, compared with 1% for Qwen3.8-27B, which indicates whether larger budgets would convert failures into passes.

Primary failure reasons. Each non-`PASS` episode is assigned exactly one primary reason, determined from its scorer gates, termination, rejected-action share (a malformed-output loop has at least 50% rejected actions), and wall time; no case was adjudicated by hand. In the reference-harness condition of the renal L6 recovery task (five agents, ten subjects, one seed, 50 episodes), the counts are: exhausting the budget without a correction: 16 (32%); declaring done without a correction: 12 (24%); a malformed-output loop: 9 (18%); reacquiring without changing the responsible setting: 4 (8%); abstaining on a correctable case: 2; never declaring done on a clean case: 2; changing the responsible setting with coverage still incomplete: 1. The remaining 4 episodes passed. Claude Opus 5.5 adds a further pattern, over-correction: it altered a correct acquisition in all 10 clean L4 and L6 cells and abstained on all 30 clean L2 cells. The dominant failure is thus not a lack of observation but completing the workflow without changing the responsible acquisition setting.

The supplementary material provides annotated CT and MR trajectories that align visible observations and actions with workflow and acquisition-quality gates.

7 Discussion, Future Directions, and Conclusion

Future directions. Future work may study adaptation from deidentified operator traces and reviewable assistance that can expose actions, pause before acquisition, and retain trajectories. Deployment requires appropriate privacy, security, and regulatory safeguards (U.S. Department of Health and Human Services, 2026); current episodes contain no human correction.

Scope and limitations. The benchmark represents selected procedures through simulated consoles and task-specific physical abstractions. Its geometry and coverage measurements characterize acquired anatomy under these models; clinical image quality and diagnostic utility require additional evaluation. The ladder organizes task requirements, although observed failures may involve interacting abilities. Results depend on the subject cohort, task versions, and agent interface; broader anatomy, consoles, and acquisition conditions remain to be studied.

Conclusion. MedImageOSWorld connects graphical console interaction with the acquisition of medical image evidence across simulated CT, MR, and ultrasound consoles. Agents configure and plan acquisitions, inspect images, and act on feedback. Task-specific workflow checks and anatomical ground truth distinguish procedural progress from acquisition of the intended anatomy, supporting comparison of acquisition agents and analysis of unsuccessful decisions. Current agents operate the consoles far better than they acquire the intended anatomy: the best open-weight agent reaches three quarters of the workflow checkpoints but succeeds on only a quarter of the weighted tasks, and no agent succeeds reliably at closed-loop control.

Ethics Statement

Benchmark episodes use virtual subjects in simulated imaging consoles. They do not acquire scans from animals or humans.

Reproducibility Statement

The supplementary material documents the experimental tasks and episodes, the subjects and assets used, simulator and scorer versions, and agent configurations. It also provides prompts, action formats, rules for accessing reference information, random seed settings, criteria for excluding or repeating attempts, stopping rules, and the scripts used to generate tables and figures. Interaction logs connect the reported results to the observations and actions recorded during each episode. File checksums and access instructions, including those for licensed assets and restricted scoring components, have yet to be finalized.

AI Use Statement

Generative AI tools assisted with manuscript drafting and revision, literature searches, and the interpretation and organization of experimental reports. AI coding agents, including Claude Code with Opus 5 and Fable 5.1 and Codex CLI with GPT-5.6 Sol and GPT-6 Astra, assisted with implementing test environment, orchestrating cluster runs, developing scripts for analysis and reporting, and drafting findings based on per-episode test logs and groundtruth. An AI model assigned the final VLM scores from these logs according to scoring rules made by human experts. The authors reviewed and tested the AI-generated code, examined the experimental outputs, and checked the reported numerical results against the underlying experimental data. The authors take responsibility for the final text, claims, and artifacts.

References

José M. Algarín, Teresa Guallart-Naval, José Borreguero, Fernando Galve, and Joseba Alonso. MaRGE: A graphical environment for MaRCoS. *Journal of Magnetic Resonance*, 361:107662, 2024. doi: 10.1016/j.jmr.2024.107662. URL https://doi.org/10.1016/j.jmr.2024.107662.

Suhana Bedi, Ryan Welch, Ethan Steinberg, Michael Wornow, Taeil Matthew Kim, Haroun Ahmed, Peter Sterling, Bravim Purohit, Qurat Akram, Angelic Acosta, Esther Nubla, Pritika Sharma, Michael A. Pfeffer, Sanmi Koyejo, and Nigam H. Shah. HealthAdminBench: Evaluating computer-use agents on healthcare administration tasks. In *ICLR 2026 Workshop on Agents in the Wild (AIWILD)*, 2026. URL https://openreview.net/forum?id=NNLD7776OF. Non-archival workshop paper.

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://doi.org/10.1056/AIdbp2500144.

Bulat Maksudov, Vladislav Kurenkov, Kathleen M. Curran, and Alessandra Mileo. ABRA: Agent benchmark for radiology applications. arXiv preprint arXiv:2605.11224v1, 2026. URL https://arxiv.org/abs/2605.11224v1. Version 1, 2026-05-11.

Keerthi Sravan Ravi and Sairam Geethanath. Autonomous magnetic resonance imaging. *Magnetic Resonance Imaging*, 73:177–185, 2020. doi: 10.1016/j.mri.2020.08.010. URL https://doi.org/10.1016/j.mri.2020.08.010.

Chris Rawles, Sarah Clinckemaillie, Yifan Chang, Jonathan Waltz, Gabrielle Lau, Marybeth Fair, Alice Li, William Bishop, Wei Li, Folawiyo Campbell-Ajala, Daniel Toyama, Robert Berry, Divya Tyamagundlu, Timothy Lillicrap, and Oriana Riva. AndroidWorld: A dynamic benchmarking environment for autonomous agents. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/01a83bc2f2732a58e6aa731e659e7101-Abstract-Conference.html.

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*, 9:499, 2026. doi: 10.1038/s41746-026-02674-7. URL https://doi.org/10.1038/s41746-026-02674-7.

W. P. Segars, G. Sturgeon, S. Mendonca, J. Grimes, and B. M. W. Tsui. 4d XCAT phantom for multimodality imaging research. *Medical Physics*, 37(9):4902–4915, 2010. doi: 10.1118/1.3480985. URL https://doi.org/10.1118/1.3480985.

Rozain Shakeel, Abdul Rahman Mohammad Ali, Muneeb Mushtaq, Tausifa Jan Saleem, and Tajamul Ashraf. MedSPOT: A workflow-aware sequential grounding benchmark for clinical GUI. In Paolo Favaro, Zuzana Kukelova, Atsuto Maki, Anna Rohrbach, Konrad Schindler, and Federico Tombari (eds.), *Computer Vision – ECCV 2026*, volume 17066 of *Lecture Notes in Computer Science*, pp. 240–257. Springer Nature Switzerland, 2026. doi: 10.1007/978-3-032-37718-0_15. URL https://doi.org/10.1007/978-3-032-37718-0_15.

U.S. Department of Health and Human Services. Security standards for the protection of electronic protected health information. 45 C.F.R. Part 164, Subpart C, 2026. URL https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-C/part-164/subpart-C. eCFR text current as of September 24, 2026; accessed September 26, 2026.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1650. URL https://proceedings.nips.cc/paper_files/paper/2024/hash/5d413e48f84dc61244b6be550f1cd8f5-Abstract-Datasets_and_Benchmarks_Track.html.

Jia Yu, Zilong Wang, Xinyang Jiang, Dongsheng Li, and Shuo Wang. MedCUA-Bench: A screenshot-only benchmark for clinical computer-use agents. arXiv preprint arXiv:2606.03203v1, 2026. URL https://arxiv.org/abs/2606.03203v1. Version 1, 2026-06-02.

Mengqi Yuan, Zilong Zhou, Xinzhuang Xiong, Weiming Wu, Jiayang Sun, Jiamin Song, Kaiqian Cui, Bowen Wang, Haoyuan Wu, Yitong Li, Dunjie Lu, Haikong Lu, Qi Zhen, Xinyuan Wang, Jiaqi Deng, Yuhao Yang, Cheng Chen, Boyuan Zheng, Alex Su, Xiao Yu, Hao Zou, Saaket Agashe, Xing Han Lu, Manpreet Kaur, Zhengyang Qi, Vincent Sunn Chen, Frederic Sala, Dayiheng Liu, Junyang Lin, Zhou Yu, Yu Su, Siva Reddy, Xin Eric Wang, Peng Qi, Tianbao Xie, and Tao Yu. OSWorld 2.0: Benchmarking computer use agents on long-horizon real-world tasks. arXiv preprint arXiv:2606.29537v2, 2026. URL https://arxiv.org/abs/2606.29537v2. Version 2, 2026-07-13.

Qiaoyu Zheng, Chaoyi Wu, Weike Zhao, Pengcheng Qiu, Lisong Dai, Ya Zhang, Yanfeng Wang, and Weidi Xie. How well can modern LLMs act as agent cores in radiology environments? arXiv preprint arXiv:2412.09529v4, 2024. URL https://arxiv.org/abs/2412.09529v4. Version 4, 2026-09-07.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/4410c0711e9154a7a2d26f9b3816d1ef-Abstract-Conference.html.