Watch Once, Work Forever: One-Shot Computer-Use Agents for Clinical UI Automation

Udbhav S. Ram[1,2], Carlos E. Cardenas[1], and Rex A. Cardan[1]

1. Department of Radiation Oncology, The University of Alabama at Birmingham, Birmingham, Alabama, USA
2. Department of Physics and Astronomy, McMaster University, Hamilton, Ontario, Canada

2026 JOINT AAPM | COMP MEETING
JULY 19–22 | VANCOUVER, BC

AIMS

Radiotherapy planning requires frequent interaction with complex clinical software, creating inefficiency and opportunities for error. Current automation pipelines depend on scripting, large datasets, or vendor-specific APIs. Here, we assess whether a one-shot, pixel-to-action computer-use agent can learn a workflow from a single demonstration and generalize across interface states.

METHODS

A computer use agent (CUA) was developed to perform common tasks within a sandbox environment, including plan navigation, structure review, DVH extraction, and report generation in the Eclipse TPS. The agent leverages a multimodal perception-action model combining visual interface understanding, natural-language task representation, and one-shot learning from a single un-annotated demonstration.

After observation, the agent was evaluated on unseen tasks and workflow variations without additional fine-tuning. Performance was assessed by task completion rate, execution time, and agreement between agent-generated and human-generated ground truth outputs. Failure behaviours were also analyzed to assess clinical feasibility.

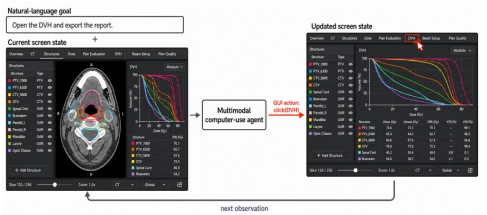

Figure 1: Sample image + natural language input and corresponding grounding model output (FARA-7B)

RESULTS

- Candidate vision–language backbones were benchmarked on a 100-prompt UI localization suite (53 coordinate click tasks, 47 boolean UI-state questions; temperature=0).
- Qwen3-VL-8B led spatial grounding (79.2% clicks within 50 px; 12.5 px median error; 25.23 s/prompt) while maintaining 100% boolean accuracy.
- Qwen3-VL-4B offered the best speed–accuracy tradeoff (75.5% within 50 px; 16.0 px median error; 15.29 s/prompt; 97.9% boolean).
- UI-TARS-7B-DPO was similar in localization (73.6% within 50 px; 15.3 px median error) with 100% boolean accuracy but higher latency (31.09 s/prompt).
- FARA variants retained strong semantic performance (97.9–100% boolean) yet failed at spatial grounding (0% within 50 px; ~700 px median error), revealing a semantic–spatial split that limits reliable autonomous clicking in complex clinical GUIs.

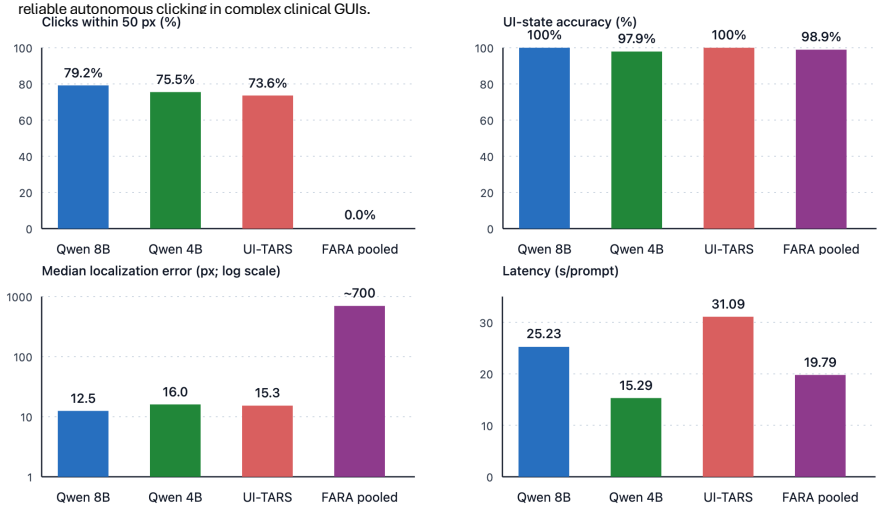

Figure 2: Performance of candidate vision–language backbones on the 100-prompt clinical UI benchmark (53 coordinate-click tasks and 47 UI-state questions; temperature = 0).

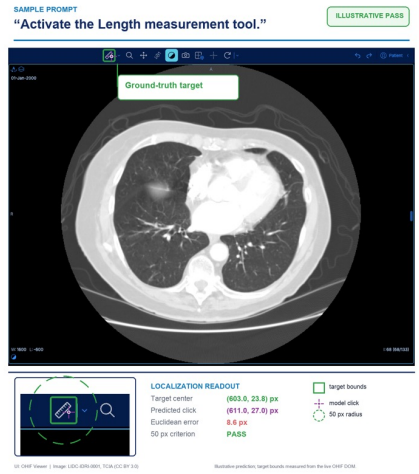

Figure 3: Clinical UI localization example (Qwen 8B)

CONCLUSIONS

One-shot learning computer use agents demonstrate promising capability to autonomously execute complex dosimetric workflow tasks after minimal supervision. These findings support further investigation into agent-assisted treatment planning as a human-in-the-loop clinical tool.

CONTACT INFORMATION

rcardan@uabmc.edu @rexcardan
cecardenas@uabmc.edu @carlosmedphys
Udbhavram41@gmail.com @UdbhavRam

Abstract Info