MedSPOT: A Workflow-Aware Sequential Grounding Benchmark for Clinical GUI

Rozain Shakeel[1], Abdul Rahman Mohammad Ali[2], Muneeb Mushtaq[1], Tausifa Jan Saleem[3], and Tajamul Ashraf[1,4⋆]

[1] Gaash Research Lab, National Institute of Technology Srinagar, India
[2] e& Group, UAE
[3] Mohammad Bin Zayed University of Artificial Intelligence (MBZUAI), UAE
[4] King Abdullah University of Science and Technology (KAUST), Saudi Arabia

Project Page: https://rozainmalik.github.io/MedSPOT_web/

Abstract. Despite the rapid progress of Multimodal Large Language Models (MLLMs), their ability to perform reliable visual grounding in high-stakes clinical software environments remains underexplored. Existing GUI benchmarks largely focus on isolated, single-step grounding queries, overlooking the sequential, workflow-driven reasoning required in real-world medical interfaces, where tasks evolve across interdependent steps and dynamic interface states. We introduce **MedSPOT**, a workflow-aware sequential grounding benchmark for clinical GUI environments. Unlike prior benchmarks that treat grounding as a standalone prediction task, **MedSPOT** models procedural interaction as a sequence of structured spatial decisions. The benchmark comprises 216 task-driven videos with 597 annotated keyframes, in which each task comprises 2–3 interdependent grounding steps within realistic medical workflows. This design captures interface hierarchies, contextual dependencies, and fine-grained spatial precision under evolving conditions. To evaluate procedural robustness, we propose a strict sequential evaluation protocol that terminates task assessment upon the first incorrect grounding prediction, explicitly measuring error propagation in multi-step workflows. We further introduce a comprehensive failure taxonomy, including edge bias, small-target errors, no prediction, near miss, far miss, and toolbar confusion, to enable systematic diagnosis of model behavior in clinical GUI settings. By shifting evaluation from isolated grounding to workflow-aware sequential reasoning, **MedSPOT** establishes a realistic and safety-critical benchmark for assessing multimodal models in medical software environments. Code and data are available at https://github.com/Tajamul21/MedSPOT

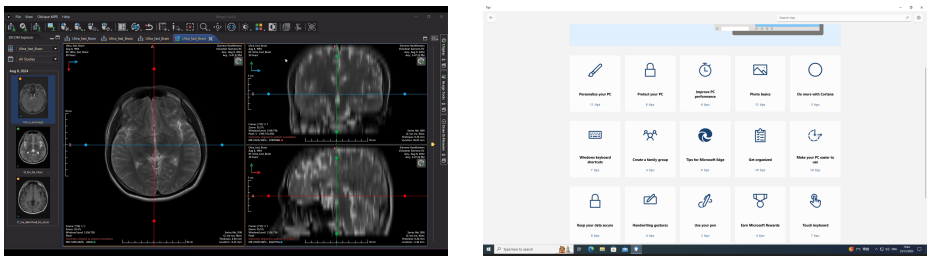
1 Introduction

Graphical User Interfaces (GUIs) are the primary medium through which users interact with modern software systems. *GUI grounding* refers to the task of mapping natural language instructions to precise visual and interactive elements

⋆ Corresponding author: tajamul21.ashraf@gmail.com

Table 1: Comparison of **MedSPOT** against existing GUI grounding benchmarks. Green cells indicate the presence of a capability, red cells indicate absence. **MedSPOT** is the only benchmark that jointly evaluates workflow-aware sequential grounding, early-termination protocol, structured failure taxonomy, and medical software environments.

Benchmark	Size	Multi-Step Tasks	Step Eval.	Sequential Protocol	Failure Taxonomy	Medical Software
ScreenSpot [7]	1,272	×	×	×	×	×
ScreenSpot-Pro [20]	1,581	×	✓	×	×	×
OSWorld-G [48]	564	✓	×	×	×	×
OmniACT [18]	9,802	✓	✓	×	×	×
AssistGUI [11]	100	✓	✓	×	×	×
Mind2Web [8]	2,350	✓	×	×	×	×
MedSPOT (Ours)	**216 / 597**	✓	✓	✓	✓	✓



(a) Medical GUI interface **(b)** General GUI interface

Fig. 1: Comparison of interface complexity. (a) Medical GUI environments are densely structured, hierarchically organized, and semantically rich with domain-specific terminology. (b) General GUI environments typically involve simpler layouts and short-horizon navigation tasks.

within an interface [7, 45]. Recent advances in Vision-Language Models (VLMs) and Multimodal Large Language Models (MLLMs) [3, 23, 35, 38, 39] have significantly improved multimodal reasoning capabilities, enabling systems to interpret screenshots, detect interface elements, reason over contextual cues, and predict executable actions.

Despite this progress, existing research predominantly focuses on general-purpose environments such as web navigation and desktop automation. Benchmarks such as ScreenSpot [7], Mind2Web [8], OmniAct [18], and related works [2, 12, 20, 54] evaluate grounding in short-horizon, single-step scenarios. These settings treat grounding as an isolated prediction problem, overlooking the structured, sequential reasoning required in real-world workflows. This limitation becomes critical in high-stakes domains such as clinical software systems. Modern medical workflows rely heavily on complex graphical interfaces, including diagnostic imaging viewers, treatment planning systems, and health record dashboards. As illustrated in Fig. 1, medical GUIs are densely structured, hierarchically organized, and semantically rich with domain-specific terminology. Within

such environments, accurate grounding requires not only visual alignment but also contextual awareness of interface hierarchies, multi-step procedural reasoning, and strict precision. An early grounding error can invalidate subsequent actions, potentially leading to severe operational consequences.

Yet, despite the rapid growth of both GUI grounding and medical AI research, there exists no benchmark specifically designed to evaluate *workflow-aware sequential grounding* in clinical software environments. Existing GUI datasets focus on general-purpose settings, while medical benchmarks [5, 15, 51] primarily evaluate static image reasoning rather than interactive procedural workflows. As a result, the intersection of sequential grounding and safety-critical medical interfaces remains largely unexplored. As summarized in Table 1, existing GUI grounding benchmarks primarily focus on general-purpose environments and short-horizon tasks. While several datasets support multi-step interactions, none incorporate a strict sequential evaluation protocol, structured failure analysis, or domain-specific clinical software settings. In contrast, **MedSPOT** is the first benchmark to jointly evaluate workflow-aware grounding under early-termination scoring in safety-critical medical interfaces.

To address this gap, we introduce **MedSPOT**, a benchmark for evaluating workflow-aware spatial grounding in clinical GUI environments. **MedSPOT** spans **10 functionally diverse medical imaging software platforms** and consists of **216 task-driven videos with 597 annotated keyframes**, as illustrated in Fig. 2. Each task involves multiple interdependent grounding steps that reflect realistic clinical procedures, where decisions evolve across dynamic interface states. Beyond dataset construction, we propose a *failure-aware sequential evaluation protocol* that terminates task scoring upon the first incorrect grounding prediction. This strict early-termination strategy explicitly measures error propagation across multi-step workflows, providing a more realistic assessment of procedural robustness. Furthermore, we introduce a structured six-class failure taxonomy, including edge bias, small-target errors, no prediction, near miss, far miss, and toolbar confusion, to systematically diagnose grounding behavior in complex medical interfaces. More details regarding MedSPOT are given in Supplementary Material A.

Our contributions are summarized as follows:

- We introduce **MedSPOT**, the first benchmark for workflow-aware sequential grounding in clinical GUI environments, spanning 10 medical software platforms with 216 tasks and 597 annotated keyframes.
- We propose a *failure-aware sequential evaluation protocol* with early termination, capturing realistic error propagation in safety-critical multi-step workflows.
- We define a structured *six-class failure taxonomy* to systematically analyze grounding errors in densely structured medical interfaces.
- We conduct a comprehensive evaluation of 16 state-of-the-art MLLMs, revealing a substantial performance gap between general-purpose models and GUI-specialized architectures in clinical environments.

2 Related Work

2.1 GUI Grounding Benchmarks and Agent Models

Recent benchmarks have advanced GUI grounding in general-purpose environments. ScreenSpot [7] introduced a cross-platform grounding dataset across mobile, desktop, and web interfaces, while ScreenSpot-Pro [20] extended evaluation to high-resolution professional applications, highlighting degradation on small, dense UI elements. VenusBench-GD [55] and MMBench-GUI [42] further expanded coverage with hierarchical evaluation protocols, and web-centric datasets such as Mind2Web [8] and OmniACT [18] incorporated multi-step browsing tasks. Parallel to benchmark development, GUI grounding models have progressed rapidly. UI-TARS [30] integrates perception, unified action modeling, and System-2 reasoning at scale; Agent-X [2] emphasizes structured multi-step reasoning; and GUI-Actor [45] proposes coordinate-free grounding via a dedicated <ACTOR> token. General-purpose backbones such as Qwen2.5-VL [38], Qwen3-VL [39], and GPT-style models remain strong baselines. Despite this progress, existing benchmarks focus on general-purpose software, evaluate largely independent steps without strict inter-step dependency, and lack early-termination protocols to measure error propagation. As summarized in Table 1, none jointly address sequential dependency, structured failure analysis, and clinical software environments. **MedSPOT** fills this gap by modeling grounding as interdependent spatial decisions under a strict sequential evaluation protocol. Desktop GUI grounding has been explored in [9, 29]. Concurrent work Chain-of-Ground [19] improves single-step grounding via iterative reasoning, unlike MedSPOT's multi-step sequential evaluation.

2.2 Medical Benchmarks and Clinical Software Environments

Medical AI benchmarks have predominantly focused on static image understanding and clinical question answering [1]. Datasets such as MedSG-Bench [51], GMAI-MMBench [5], and OmniMedVQA [15] evaluate multimodal reasoning over radiological images and medical reports. VividMed [22] advances medical visual grounding but operates on static images. While these works advance domain-specific reasoning, they do not consider interactive GUI-based workflows or spatial grounding within software interfaces. In practice, modern clinical workflows rely heavily on specialized software platforms including Orthanc [17], Weasis [33], RadiAnt [25], BlueLight DICOM Viewer [6], MicroDICOM [26], 3D Slicer [10], ITK-SNAP [52], MITK [43], Ginkgo CADx [44], and DICOMscope [27]. These systems feature densely structured toolbars, hierarchical menus, domain-specific terminology, and fine-grained interaction mechanisms that impose unique grounding challenges not captured in existing GUI datasets. To our knowledge, **MedSPOT** is the first benchmark to systematically evaluate workflow-aware sequential grounding within real-world clinical GUI environments, bridging the gap between multimodal reasoning research and safety-critical medical software interaction.

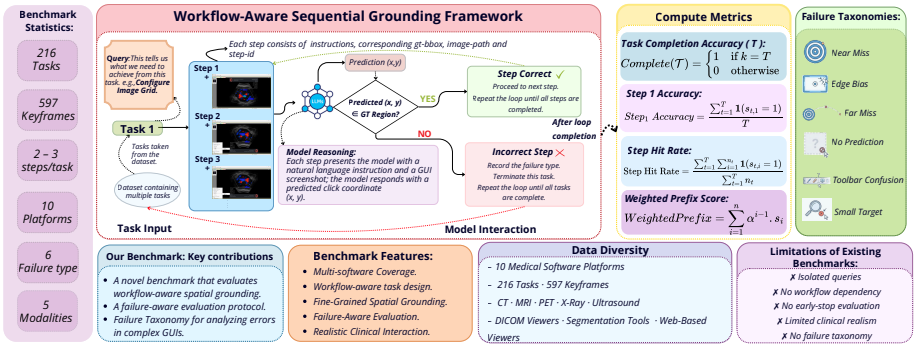


Fig. 2: Overview of the **MedSPOT** dataset construction and sequential evaluation pipeline.

3 MedSPOT Benchmark and Evaluation

Problem Formulation and Dataset Design. We formulate workflow-aware `GUI` grounding in clinical software as a *sequential, instruction-conditioned spatial localization problem*. Unlike conventional grounding benchmarks that evaluate isolated predictions, **MedSPOT** models grounding as a temporally dependent sequence of spatial decisions within evolving interface states.

Let $\mathcal{S}$ denote the set of medical software platforms. The dataset is defined as: $\mathcal{D} = \bigcup_{s \in \mathcal{S}} \mathcal{D}_s$, where $\mathcal{D}_s$ represents the task set associated with software s. Each $\mathcal{D}_s$ consists of N_s task workflows: $\mathcal{D}_s = \{\mathcal{T}_i\}_{i=1}^{N_s}$. Each task $\mathcal{T}$ is a finite ordered sequence of T interaction steps: $\mathcal{T} = \{(I_t, s_t, A_t)\}_{t=1}^{T}$, where:

- $I_t \in \mathbb{R}^{H \times W \times 3}$ denotes the `GUI` frame at step t,
- s_t is the natural language instruction,
- A_t is the corresponding ground-truth action.

Each action is defined as: $A_t = (a_t, y_t, B_t)$, where:

- $a_t \in \mathcal{A} = \{\texttt{click}\}$,
- y_t is the semantic target description,
- $B_t = (x, y, w, h) \in [0, 100]^4$ is the normalized bounding box in percentage coordinates.

Importantly, the steps within $\mathcal{T}$ are *interdependent*. An incorrect grounding at time t invalidates downstream actions, reflecting the procedural dependency structure inherent in clinical workflows. This sequential dependency differentiates **MedSPOT** from prior `GUI` grounding datasets that treat each instruction-frame pair independently.

3.1 Annotation Protocol

For each medical software platform $s \in \mathcal{S}$, we record real `GUI` interaction workflows as video sequences: $\mathcal{V}_i = \{F_1, F_2, \ldots, F_n\}, \quad n \geq 2$, where each action

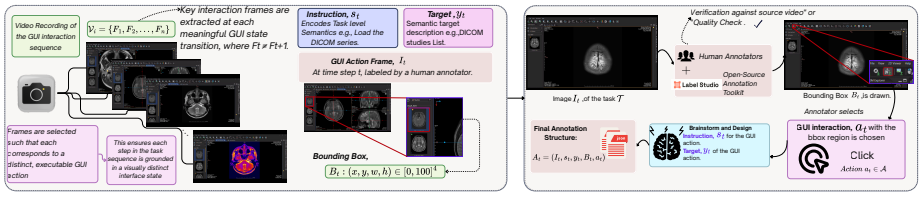


Fig. 3: Overview of the **MedSPOT** annotation pipeline. GUI interaction videos are recorded from real clinical software, segmented into causally consistent decision frames, and annotated with step-level instructions, semantic targets, bounding boxes, and action types to construct sequential grounding tasks.

induces a visible GUI state transition: $F_t \neq F_{t+1}$. These sequences simulate realistic clinical workflows, including loading DICOM studies, navigating image series, applying transformations, performing measurements, and exporting results. As illustrated in Fig. 3, the annotation process converts raw clinical GUI interaction videos into temporally ordered, step-level grounding tasks through structured segmentation and multi-level labeling.

Step Extraction. From each video $\mathcal{V}_i$, we extract a minimal ordered subset of decision frames: $\{I_t\}_{t=1}^{T} \subset \mathcal{V}_i$, where each I_t corresponds to a meaningful interaction decision point and preserves temporal order. This produces a compact sequence capturing causally consistent state transitions.

Frame-Level Annotation. Each frame is annotated using Label Studio [40] as: $\mathcal{A}_t = (I_t, s_t, y_t, B_t, a_t)$, where:

- I_t is the GUI frame encoding layout structure, widgets, modality views, and domain-specific elements,
- s_t is the natural language instruction,
- y_t is the semantic description of the target UI element,
- $B_t = (x, y, w, h) \in [0, 100]^4$ is the normalized bounding box,
- $a_t \in \mathcal{A} = \{\texttt{click}\}$.

This annotation enforces joint learning of: $P(\text{region} \mid s_t, I_t)$ and $P(y_t \mid I_t)$, linking semantic reasoning and spatial localization. Inter-annotator agreement details are reported in the Supplementary material I.

Task Construction. Each annotated interaction step is represented as $\text{step}_t = (\text{id}_t, I_t, s_t, A_t)$, where id_t denotes the step index, I_t is the GUI frame, s_t is the natural language instruction, and A_t is the corresponding ground-truth action. The steps are ordered temporally as $\text{step}_1 < \text{step}_2 < \cdots < \text{step}_T$, forming a causally consistent interaction trajectory in which the next interface state depends on the previous state and executed action, i.e., $P(I_{t+1} \mid I_t, a_t)$. Each recorded GUI interaction video produces a task $\mathcal{T}_i = \{\text{step}_1, \ldots, \text{step}_T\}$, representing a complete procedural workflow. Tasks are grouped per software platform

as $\mathcal{D}_s = \{\mathcal{T}_1, \ldots, \mathcal{T}_{N_s}\}$, with disjoint task sets such that $\mathcal{T}_i \cap \mathcal{T}_j = \emptyset$ for $i \neq j$. This hierarchical construction (step → task → software → dataset) ensures completeness, non-overlapping workflows, and preservation of causal structure. Overall, **MedSPOT** jointly evaluates spatial grounding and procedural reasoning within structured medical `GUI` workflows.

3.2 Evaluation Pipeline

Given a `GUI` frame I_t, instruction s_t, and interaction history $\mathcal{H}_t$, a model predicts click coordinates $\hat{p}_t \sim P(p \mid I_t, s_t, \mathcal{H}_t)$. Models may output a single point or a set of top-K candidates $\hat{P}_t = \{\hat{p}_{t,k}\}_{k=1}^{K}$, with $\hat{p}_{t,k} \in \mathbb{R}^2$. Ground-truth annotations are provided as normalized bounding boxes $(x, y, w, h) \in [0, 100]^4$, which are converted into pixel space as $B_t = (x_1, y_1, x_2, y_2)$ using the image dimensions (W, H). A prediction is considered correct if any predicted point lies within B_t, i.e., $x_1 \leq \hat{x} \leq x_2$ and $y_1 \leq \hat{y} \leq y_2$. For models operating on discrete visual patches, a tolerance region $\hat{B}_p = (\hat{x} \pm \delta, \hat{y} \pm \delta)$ is introduced, and a hit is accepted if $\hat{B}_p \cap B_t \neq \emptyset$.

To analyze spatial difficulty, we define the normalized target area $A(B) = \frac{(x_2 - x_1)(y_2 - y_1)}{WH}$ and categorize targets as small, medium, or large based on predefined thresholds, capturing the nonlinear degradation observed for compact `UI` elements [53].

Evaluation follows a strict sequential protocol. Step correctness is defined as $\delta_t = 1$ if $\hat{p}_t \in B_t$ and 0 otherwise. We compute the longest valid prefix $k = \max\{t \mid \forall i \leq t, \delta_i = 1\}$ and terminate evaluation upon the first failure. A task is considered complete only if $k = T$, meaning all steps are correct. This corresponds to maximizing the joint probability $\prod_{t=1}^{T} P(\hat{p}_t \in B_t \mid \mathcal{H}_t)$, implying that if per-step accuracy is p, the probability of full task completion scales as p^T. Consequently, the protocol emphasizes temporal consistency and penalizes early errors, transforming evaluation from independent step accuracy into a measure of causally consistent, workflow-aware grounding. Further details are in the Supplementary material C.

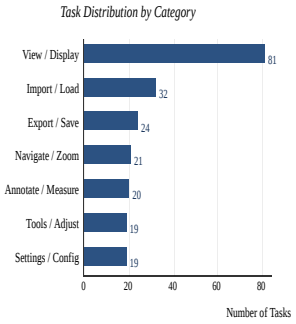
3.3 Domain Coverage and Interface Diversity

MedSPOT spans 10 open-source medical imaging platforms, covering three primary interface categories: (1) `DICOM/PACS` viewers, (2) segmentation and research tools, and (3) web-based viewers. The benchmark comprises **216** video tasks, **597** annotated keyframes, with an average of *2–3 interdependent steps per task*. The dataset captures heterogeneous `GUI` layouts, modality-specific visualization structures, hierarchical toolbars, nested menus, and diverse interaction paradigms. This diversity ensures evaluation of cross-interface generalization rather than overfitting to a single visual layout.

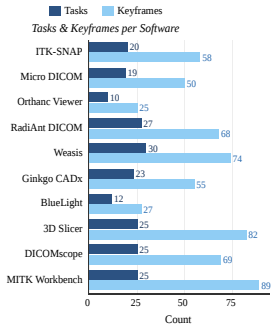
Software Coverage. The platforms include BlueLight DICOM Viewer [6], 3D Slicer [10], Ginkgo-CADx [44], DICOMscope [27], Orthanc [17], RadiAnt [25], MicroDICOM [26], Weasis [33], MITK [43], and ITK-SNAP [52]. These systems

Statistic	Value
Software Platforms	10
Total Tasks	216
Total Annotated Keyframes	597
Avg. Steps per Task	2–3
Desktop Applications	8
Web-Based Viewers	2
DICOM / PACS Viewers	109 tasks
Segmentation / Research	70 tasks
Web-Based Viewers	37 tasks
Modalities	CT, MRI, PET, X-Ray, US

(a) Dataset statistics



(b) Task categories



(c) Per-software task depth

Fig. 4: MedSPOT dataset statistics illustrating overall scale, task category distribution, and per-software coverage across ten clinical platforms.

support diverse imaging modalities, including `CT`, `MRI`, `PET`, X-ray, and Ultrasound.

Task Category Distribution. The overall dataset composition is summarized in Fig. 4. Specifically, Fig. 4a reports the global benchmark statistics, including the total number of video tasks, annotated keyframes, and average steps per task. Figure 4b illustrates the distribution of tasks across seven functional categories. View/Display operations dominate (81 tasks), reflecting the central role of visualization in clinical imaging workflows. Additional categories include Import/Load (32), Export/Save (24), Navigate/Zoom (21), Annotate/Measure (20), Tools/Adjust (19), and Settings/Configuration (19), collectively covering the procedural spectrum encountered in real clinical usage. Finally, Fig. 4c presents the per-software task and keyframe distribution, highlighting variations in interaction depth and interface complexity across platforms. Details regarding data provenance and ethics are mentioned in the supplementary D.

4 Experiments

Evaluation Protocol We employ two complementary evaluation functions: *Step-wise Evaluation.* At each step t, a prediction is considered correct if the predicted click location lies within (or sufficiently overlaps) the ground-truth bounding box. Evaluation follows an early-termination strategy: once a step is incorrect, all subsequent steps in the task are invalidated. This mirrors real-world `GUI` usage, where an incorrect interaction alters the interface state and disrupts the intended procedural trajectory. *Task Completion.* A task is considered successfully completed only if all interdependent steps are executed correctly. This metric provides a strict measure of procedural robustness and long-horizon reasoning capability.

Metrics We report the following metrics to capture both spatial precision and sequential consistency: *Task Completion Accuracy (`TCA`).* The fraction of tasks

Table 2: Comparison of state-of-the-art MLLMs on **MedSPOT-Bench** under the strict sequential evaluation protocol. We report Step Hit Rate (SHR), Step-1 Accuracy (S1A), Weighted Prefix Score (WPS), and Task Completion Accuracy (TCA), where higher values indicate stronger spatial precision and sustained workflow consistency. **Best results** are highlighted in bold green and second-best results are underlined in blue.

Model Name	Params	SHR(%)↑	S1A(%)↑	WPS↑	TCA(%)↑
Closed Source Models					
GPT 4o mini [28]	-	4.91	5.14	0.051	0.0
GPT 5 [35]	-	16.4	16.5	0.185	2.8
Open Source Models					
Llama 3.2 Vision-Instruct [13]	11B	0.0	0.0	0.0	0.0
Qwen2-VL-Instruct [41]	7B	1.83	1.87	0.02	0.0
DeepSeek-VL2 [47]	16B	2.7	2.8	0.03	0.0
Gemma 3 [37]	27B	3.7	1.3	0.04	0.0
Mistral-3-Instruct [21]	8B	6.14	6.0	0.064	0.0
UGround-V1 [12]	7B	10.17	8.88	0.107	0.93
Seeclick [7]	-	4.3	9.3	0.11	1.4
Qwen2.5-VL-Instruct [38]	7B	19.3	33.17	0.48	12.6
CogAgent [14]	9B	37.8	27.1	0.45	15.4
Aguvis [50]	7B	54.8	44	0.75	26.7
Os-Atlas [46]	7B	55	45	0.8	26.6
UI-Tars 1.5-VL [30]	7B	**66**	57.5	1.0	30.8
Qwen3-VL-Instruct [39]	8B	46.6	63.0	1.1	35.0
GUI-Actor [45]	7B	49.6	**65.0**	**1.2**	**43.5**

for which all steps are correctly executed. This is the most stringent measure of workflow reliability. *Step-1 Accuracy (S1A).* The fraction of tasks in which the first step is correctly grounded. This isolates initial perception and instruction alignment without sequential compounding effects. *Step Hit Rate (SHR).* The ratio of correctly completed steps before the first failure to the total number of evaluable steps across all tasks. This metric quantifies sequential progress under early stopping. *Weighted Prefix Score (WPS).* To emphasize the importance of early correctness, we introduce a weighted prefix formulation: $\mathtt{WPS} = \sum_{i=1}^{n} \alpha^{i-1} s_i$, where $s_i = 1$ if step i is correct and 0 otherwise, and $\alpha = 0.8$ is an exponential decay factor. This score reflects the disproportionate impact of early errors on overall workflow reliability. Further details regarding Inference Protocol are in the Supplementary material B.

Failure Analysis. For each incorrect step, we record the failure category, according to the taxonomy described in the Supplementary material E. This enables fine-grained diagnosis of systematic weaknesses, including spatial precision errors, small-target failures, toolbar confusion, and semantic misalignment.

4.1 Benchmark Results

1. Large-Scale Comparison Across MLLMs. We conduct a comprehensive evaluation of state-of-the-art multimodal large language models (MLLMs) on MedSPOT using our strict failure-aware sequential protocol. As shown in Table 2, performance is reported across both step-level and task-level metrics, including Step Hit Rate (SHR), Step-1 Accuracy (S1A), Weighted Prefix Score (WPS), and Task Completion Accuracy (TCA). The results reveal a substantial performance gap between general-purpose vision-language models and interface-specialized architectures. Despite strong performance on conventional multimodal benchmarks, widely adopted general-purpose models struggle dramatically in workflow-aware medical GUI grounding.

2. Collapse of General-Purpose MLLMs. We observe near-complete failure among several widely used multimodal models. Llama 3.2 Vision-Instruct [13], Qwen2-VL-Instruct [41], DeepSeek-VL2 [47], and Gemma 3 [37] all achieve 0% Task Completion Accuracy (TCA), despite parameter sizes ranging from 7B to 27B. Their Step Hit Rate (SHR) remains below 6%, and WPS is effectively negligible. Closed-source frontier models exhibit similar fragility. GPT-4o-mini [28] achieves only 4.91% SHR with 0% TCA, while GPT-5 [35] reaches 16.5% SHR but only 2.8% TCA. Although GPT-5 occasionally predicts the correct first step, it fails to sustain consistent multi-step execution under strict early termination. These findings indicate that parameter scale and generic multimodal pretraining do not translate into reliable spatial grounding within structured clinical interfaces. Unlike natural-image reasoning tasks, medical GUI environments demand pixel-level precision, fine-grained discrimination among visually similar toolbar elements, and hierarchical interface understanding.

3. Sequential Fragility and Error Propagation. Even stronger multimodal models that demonstrate reasonable single-step performance exhibit substantial degradation under sequential evaluation. For example, Qwen3-VL-Instruct [39] achieves 63.0% Step-1 Accuracy (S1A), yet TCA drops to 35.0%. GUI-Actor [45] achieves the highest S1A (65.0%) but completes only 43.5% of tasks. UI-Tars 1.5-VL [30] attains the highest SHR (66%) while TCA remains limited to 30.8%. This consistent drop from S1A to SHR to TCA reflects compounding error under our prefix-constrained evaluation. As visualized in Fig. 5, the performance hierarchy consistently follows: $\texttt{TCA} \ll \texttt{SHR} < \texttt{S1A}$. Even moderate per-step error rates rapidly reduce full task completion probability due to early termination and state dependency.

4. Interface-Specialized Architectures. Models explicitly designed for GUI interaction perform substantially better. Aguvis [50] and Os-Atlas [46] achieve approximately 55% SHR and 26% TCA. CogAgent [14] achieves 37.8% SHR but only 15.4% TCA, further demonstrating degradation under multi-step constraints. The strongest overall model, GUI-Actor [45], achieves 43.5% TCA (Table 2). Nevertheless, even this model fails to complete more than half of the tasks, indicating that workflow-aware reasoning in high-stakes medical GUIs remains an open challenge.

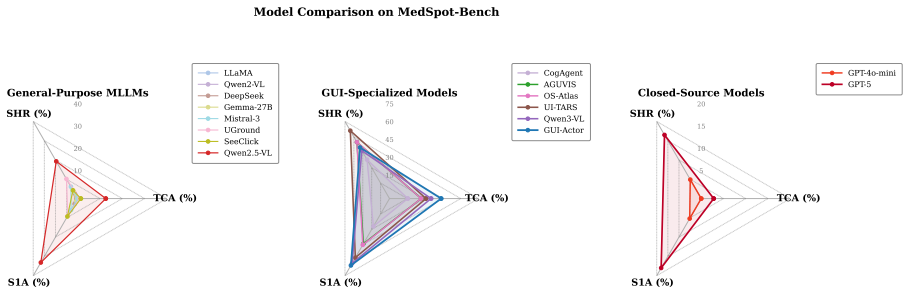


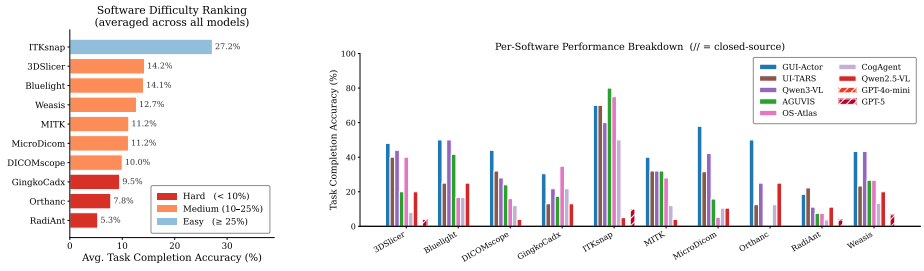
Fig. 5: Comparison of representative models across Step-1 Accuracy (`S1A`), Step Hit Rate (`SHR`), and Task Completion Accuracy (`TCA`). The consistent gap between `S1A` and `TCA` highlights compounding sequential errors under strict workflow constraints.

Table 3: Failure taxonomy distribution on MedSPOT-Bench. Open-source models are shaded in blue and closed-source models in red. Worst failure counts are highlighted in **bold red**, while lowest values are underlined blue.

Model Name	Params	Edge Bias	Far Miss	Toolbar Confusion	Near Miss	No Prediction	Small Target
Open Source Models							
Llama 3.2 Vision-Instruct [13]	11B	49	7	4	0	**134**	19
Qwen2-VL-Instruct [41]	7B	**129**	26	28	0	0	27
DeepSeek-VL2 [47]	16B	30	**152**	0	2	0	**30**
Gemma 3 [37]	27B	34	44	**110**	0	0	25
Mistral-3-Instruct [21]	8B	109	64	13	2	0	26
UGround-V1 [12]	7B	62	66	55	4	0	25
Seeclick [7]	-	103	41	21	**11**	2	29
Qwen2.5-VL-Instruct [38]	7B	20	40	78	10	10	26
CogAgent [14]	9B	57	43	36	3	16	26
Aguvis [50]	7B	42	84	31	4	0	23
Os-Atlas [46]	7B	48	49	28	5	6	21
UI-Tars 1.5-VL [30]	7B	20	55	30	0	0	22
Qwen3-VL-Instruct [39]	8B	16	35	40	6	21	18
GUI-Actor [45]	7B	26	26	21	1	24	17
Closed Source Models							
GPT 4o mini [28]	8B–10B	62	36	88	2	0	25
GPT 5 [35]	-	13	67	97	3	0	27

5. Weighted Prefix Score and Sustained Consistency. To quantify partial sequential robustness, we analyze `WPS`. As shown in Table 2, `WPS` strongly correlates with `TCA`. GUI-Actor achieves the highest `WPS` (1.2) alongside the highest `TCA` (43.5%), followed by Qwen3-VL-Instruct (1.1 `WPS`, 35.0% `TCA`) and UI-Tars (1.0 `WPS`, 30.8% `TCA`). In contrast, general-purpose `MLLMs` exhibit near-zero `WPS`, reflecting early-stage failure and minimal sustained prefix correctness. Overall, **MedSPOT** exposes a fundamental limitation of current multimodal systems: accurate first-step grounding does not imply reliable workflow execution. Robust medical `GUI` interaction requires sustained spatial precision and temporally consistent reasoning across interdependent steps.

6. Failure Taxonomy Failure analysis (Table 3) reveals distinct structural weaknesses across model families. General-purpose `MLLMs` exhibit highly concentrated failure modes: LLaMA is dominated by *No Prediction* errors (63%), DeepSeek-VL2 by *Far Miss* (71%), and Qwen2-VL and Mistral-3 by *Edge Bias*



(a) Software Difficulty Ranking

(b) Per software performance

Fig. 6: Per-software performance analysis. (a) Average `TCA` across models, ranking clinical applications by difficulty. (b) Per-model `TCA` breakdown across ten medical software platforms, highlighting interface-dependent performance variability.

(64% and 51%), indicating systematic spatial misalignment in clinical interfaces. Closed-source models show similar patterns, with `GPT-5` heavily affected by *Toolbar Confusion* (97) and *Far Miss* (67), and GPT-4o-mini exhibiting substantial *Toolbar Confusion* (88) and *Edge Bias* (62). In contrast, interface-specialized models display more distributed, difficulty-driven errors: GUI-Actor spreads failures across *Edge Bias*, *Far Miss*, and *Small Target*, while UI-TARS, Aguvis, and OS-Atlas show balanced spatial and semantic confusion patterns. Notably, *Small Target* failures persist (12–18%) across all models, reflecting the intrinsic challenge of compact toolbar icons in DICOM viewers.

4.2 Human Baseline

To assess benchmark tractability and practical utility, we conducted a preliminary human baseline study using four annotators on 45 tasks under the same screenshots, instructions, and sequential early-termination protocol used for model evaluation. As shown in Table 4, humans substantially outperform current GUI models across all evaluation metrics. These results indicate that MedSPOT remains readily solvable by human users while presenting significant challenges for existing workflow-aware grounding models, leaving considerable room for future improvement.

Table 4: Preliminary human baseline on MedSPOT. Humans substantially outperform the best model across all metrics under identical evaluation conditions.

Model	TCA (%)	S1A (%)	SHR (%)	WPS
Human Avg. (4 annotators)	**82.22**	**94.44**	**94.06**	**2.167**
GUI-Actor (best model)	43.50	65.00	49.60	1.200

4.3 Per-Software Performance Analysis

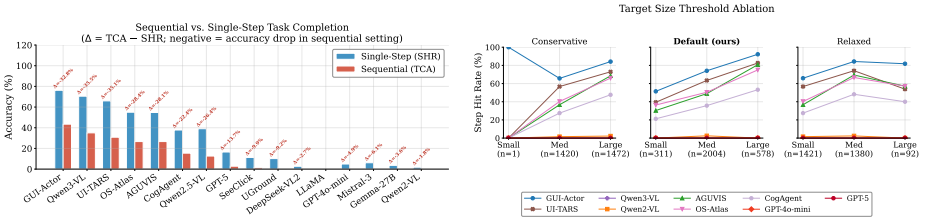
1. Software Difficulty Ranking. As shown in Fig. 6a, average Task Completion Accuracy (`TCA`) varies substantially across clinical applications, revealing strong interface-dependent performance sensitivity. RadiAnt (5.8%) and Orthanc (below 10%) emerge as the most challenging platforms, likely due to dense toolbar layouts and compact icon-heavy interfaces characteristic of DICOM viewers. Most applications fall within a moderate difficulty range (10–25%), including DICOMscope, GinkgoCADx, MITK, MicroDicom, Weasis, 3D Slicer, and BlueLight. ITKsnap stands out as the least challenging application (30.4% average `TCA`), benefiting from a comparatively structured and less cluttered interface layout. These findings confirm that layout density and interaction hierarchy significantly influence grounding reliability.
2. Per-Software Breakdown Across Models. Figure 6b presents the per-model `TCA` breakdown across all ten applications. GUI-Actor consistently achieves the highest or near-highest performance across most platforms, peaking at 70% on ITKsnap and maintaining strong results on MicroDicom (58%) and Orthanc (50%). UI-TARS follows a similar trend, also reaching 70% on ITKsnap. AGUVIS achieves the single highest score (80%) on ITKsnap but exhibits greater performance variance on denser interfaces such as RadiAnt and GinkgoCADx. While Qwen2.5-VL demonstrates moderate performance on 3D Slicer (20%) and BlueLight (25%), it collapses on DICOM-intensive tools, underscoring limited cross-interface robustness. Overall, no model exhibits uniform generalization across all software categories, reinforcing the importance of benchmark diversity and workflow-aware evaluation. Insights are mentioned in the Supplementary material F.

5 Ablation Study

We conduct systematic ablation experiments to isolate the impact of evaluation design and spatial stratification on **MedSPOT** performance. In each study, a single factor is varied while all other conditions remain fixed, enabling controlled analysis of causal effects.

5.1 Sequential vs. Single-Step Evaluation

We compare two evaluation protocols: (i) **Sequential**, where a task is considered complete only if all steps succeed in order under early termination, and (ii) **Single-Step**, where each step is scored independently without enforcing temporal dependency. As illustrated in Fig. 7a, Sequential evaluation (`TCA`) yields consistently lower scores than Single-Step Accuracy (`SHR`) across all models. The gap between `SHR` and `TCA` reflects compounding error under multi-step execution. Although some models show moderate single-step accuracy, their performance drops sharply under early termination, indicating that independent step scoring overestimates true task-level reliability.



(a) Sequential (`TCA`) vs. Single-Step (`SHR`) evaluation.

(b) Step hit rate across size tiers under varying thresholds.

Fig. 7: Ablation studies on evaluation design and spatial stratification. (a) Sequential evaluation significantly lowers task-level scores due to error compounding. (b) Small `UI` elements consistently exhibit lower accuracy across all threshold configurations.

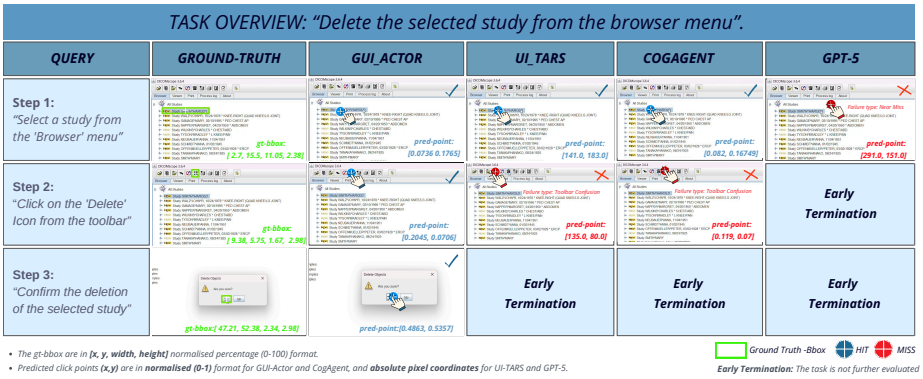


Fig. 8: Qualitative comparison on the task *"Delete the selected study from the browser menu"* (DICOMscope). GUI-Actor completes all steps, while other models fail due to toolbar confusion or near-miss predictions, resulting in early termination. Green box = ground truth; blue = correct prediction; red = incorrect prediction. Further examples are in the Supplementary material H.

5.2 Impact of Target Size Thresholds

Medical `GUI` elements range from large visualization panels to compact toolbar icons, introducing substantial variation in spatial difficulty. We stratify steps into *small* ($< 4\times10^{-4}$), *medium* (4×10^{-4}–3×10^{-3}), and *large* ($\geq 3\times10^{-3}$) targets based on normalized bounding box area, and ablate these thresholds under *Conservative*, *Default*, and *Relaxed* configurations (Fig. 7b). While the threshold choice alters category distribution (from 1 to 1,421 small targets), performance ordering remains stable across settings: small elements are consistently the hardest to localize. All models exhibit marked degradation on compact UI components prevalent in dense `DICOM` viewers and annotation tools, confirming that the difficulty is intrinsic rather than threshold-induced. Notably, Qwen2-VL approaches near-zero accuracy across size regimes, underscoring insufficient spatial precision for fine-grained medical `GUI` grounding.

5.3 Qualitative Analysis

To illustrate sequential fragility in practice, Fig. 8 presents a representative DICOMscope task. GUI-Actor successfully completes all three interaction steps, while UI-TARS and CogAgent fail at step 2 due to toolbar confusion, triggering early termination. GPT-5 fails at the first step with a near-miss prediction and cannot recover. These examples highlight a key insight: partial grounding competence does not imply workflow-level robustness under state dependency.

6 Conclusion

We introduced **MedSPOT**, the first benchmark for workflow-aware spatial grounding in clinical software interfaces. Spanning 216 task-driven videos and 597 annotated keyframes across 10 medical platforms, **MedSPOT** models grounding as a sequential, state-dependent process rather than isolated click prediction. Our strict early-termination protocol reveals substantial fragility in current multimodal models: even the strongest architecture achieves only 43.5% Task Completion Accuracy, highlighting the intrinsic difficulty of sustained grounding in dense clinical `GUIs`. Through comprehensive evaluation and failure taxonomy analysis, we demonstrate that performance degradation arises from systematic spatial and interface-structure challenges rather than incidental errors. **MedSPOT** establishes a realistic and safety-critical benchmark for multimodal interaction research in healthcare software, and we release all data and evaluation tools to support future advances in workflow-aware medical automation.

Limitations. MedSPOT is currently monolingual, limited in scale (216 video tasks), and supports only click-based interactions, excluding common actions such as drag, scroll, and text input. Evaluation is restricted to spatial grounding by checking whether predicted coordinates fall within ground-truth `GUI` bounding boxes, abstracting richer interaction semantics. Additionally, the strict early-termination protocol, while realistic, may underestimate partial procedural competence by invalidating subsequent steps after a single error. Detailed Limitations in Supplementary Material G.

References

1. Ashraf, T., Rangarajan, K., Gambhir, M., Gauba, R., Arora, C.: D-master: mask annealed transformer for unsupervised domain adaptation in breast cancer detection from mammograms. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 189–199. Springer (2024) 4
2. Ashraf, T., Saqib, A., Ghani, H., AlMahri, M., Li, Y., Ahsan, N., Nawaz, U., Lahoud, J., Cholakkal, H., Shah, M., Torr, P., Khan, F.S., Anwer, R.M., Khan, S.: Agent-x: Evaluating deep multimodal reasoning in vision-centric agentic tasks. In: ICLR (2026) 2, 4
3. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023) 2

4. Chen, G., et al.: Lion: Empowering multimodal large language model with dual-level visual knowledge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 43
5. Chen, P., Ye, J., Wang, G., Li, Y., Deng, Z., Li, W., Li, T., Duan, H., Huang, Z., Su, Y., et al.: Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. Advances in Neural Information Processing Systems **37**, 94327–94427 (2024) 3, 4
6. Chen, T.T., et al.: Bluelight: An open source dicom viewer using low-cost computation algorithm implemented with javascript using advanced medical imaging visualization. Journal of Digital Imaging **36**(2), 753–763 (2023). https://doi.org/10.1007/s10278-022-00746-0 4, 7, 21
7. Cheng, K., Sun, Q., Chu, Y., Xu, F., YanTao, L., Zhang, J., Wu, Z.: SeeClick: Harnessing GUI grounding for advanced visual GUI agents. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 9313–9332. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024), https://aclanthology.org/2024.acl-long.505 2, 4, 9, 11
8. Deng, X., Gu, Y., Zheng, B., Chen, S., Stevens, S., Wang, B., Sun, H., Su, Y.: Mind2web: Towards a generalist agent for the web. In: Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track (2023), https://proceedings.neurips.cc/paper_files/paper/2023/hash/5950bf290a1570ea401bf98882128160-Abstract-Datasets_and_Benchmarks.html 2, 4, 44
9. El Ettifouri, H., López Espejel, J., Minkova, L., Dardouri, T., Dahhane, W.: Visual grounding methods for efficient interaction with desktop graphical user interfaces. arXiv e-prints pp. arXiv–2407 (2024) 4
10. Fedorov, A., Beichel, R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J.C., Pujol, S., Bauer, C., Jennings, D., Fennessy, F., Sonka, M.: 3d slicer as an image computing platform for the quantitative imaging network. Magnetic Resonance Imaging **30**(9), 1323–1341 (2012), https://www.slicer.org 4, 7, 21
11. Gao, D., Ji, L., Zhou, Z., Wang, M., Yan, P., Zhan, D., Qian, X., Lin, J., Zhang, J., Piao, Y., Peng, Y., Zhang, Y., Jiang, Y.G., Lin, H.: Assistgui: Task-oriented desktop graphical user interface automation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19676–19686 (2024) 2
12. Gou, B., Wang, R., Zheng, B., Xie, Y., Chen, C., et al.: Navigating the digital world as humans do: Universal visual grounding for gui agents. In: ICLR (2025), https://openreview.net/forum?id=kxnoqaisCT 2, 9, 11
13. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024) 9, 10, 11
14. Hong, W., Wang, W., Lv, Q., Xu, J., Yu, W., Ji, J., Wang, Y., Wang, Z., Dong, Y., Ding, M., Tang, J.: Cogagent: A visual language model for gui agents. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14281–14290 (June 2024) 9, 10, 11
15. Hu, Y., et al.: Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 3, 4
16. Jing, H., Chen, J., Rao, C., Dang, Z., Teng, J., Chu, T., Mo, J., Fang, S., Lin, H., Lv, R., Ma, C., Zhao, L.: Sparkui-parser: Enhancing gui perception with robust grounding and parsing (2025) 42

17. Jodogne, S.: The orthanc ecosystem for medical imaging. Journal of digital imaging **31**(3), 341–352 (2018) 4, 7, 21
18. Kapoor, R., Butala, Y.P., Russak, M., Koh, J.Y., Kamble, K., AlShikh, W., Salakhutdinov, R.: Omniact: A dataset and benchmark for enabling multimodal generalist autonomous agents for desktop and web. In: Computer Vision – ECCV 2024. pp. 161–178. Springer (2024) 2, 4, 44
19. Li, A.Y., Yu, B., Lei, D., Ren, T., Liu, S.: Chain-of-ground: Improving gui grounding via iterative reasoning and reference feedback. arXiv preprint arXiv:2512.01979 (2025) 4
20. Li, K., Meng, Z., Lin, H., Luo, Z., Tian, Y., Ma, J., Huang, Z., Chua, T.S.: Screenspot-pro: Gui grounding for professional high-resolution computer use. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 8778–8786 (2025) 2, 4
21. Liu, A.H., Khandelwal, K., Subramanian, S., Jouault, V., Rastogi, A., Sadé, A., Jeffares, A., Jiang, A., Cahill, A., Gavaudan, A., et al.: Ministral 3. arXiv preprint arXiv:2601.08584 (2026) 9, 11
22. Luo, L., Tang, B., Chen, X., Han, R., Chen, T.: Vividmed: Vision language model with versatile visual grounding for medicine. In: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 1800–1821 (2025) 4
23. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 11–20 (2016) 2
24. Massih, P.A., Cosatto, E.: Reasoning with pixel-level precision: Qvlm architecture and squid dataset for quantitative geospatial analytics (2026) 41
25. Medixant: Radiant dicom viewer, https://www.radiantviewer.com 4, 7, 21
26. MicroDicom: Microdicom - free dicom viewer and software (2024), https://www.microdicom.com/, version 2024.2 4, 7, 21
27. OFFIS e.V.: DICOMscope: A free DICOM viewer. https://dicom.offis.de/dicomscope (2024) 4, 7, 21
28. OpenAI: Gpt-4o mini (2024), https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/, accessed: 2024-12-24 9, 10, 11
29. Qian, Y., Lu, Y., Hauptmann, A.G., Riva, O.: Visual grounding for user interfaces. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track). pp. 97–107 (2024) 4
30. Qin, Y., Ye, Y., Fang, J., Wang, H., Liang, S., Tian, S., Zhang, J., Li, J., Li, Y., Huang, S., et al.: Ui-tars: Pioneering automated gui interaction with native agents (2025) 4, 9, 10, 11
31. Qiu, H., Gao, P., Lu, L., Zhang, X., Shao, L., Lu, S.: Spatial preference rewarding for mllms spatial understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 720–730 (2025) 43
32. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 658–666 (2019) 49
33. Roduit, N.: Weasis dicom viewer, https://github.com/nroduit/Weasis 4, 7, 21
34. Rutherford, M., Mun, S.K., Levine, B., Bennett, W., Smith, K., Farmer, P., Jarosz, Q., Wagner, U., Freyman, J., Blake, G., et al.: A dicom dataset for evaluation of medical image de-identification. Scientific Data **8**(1), 183 (2021) 35, 46

35. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025) 2, 9, 10, 11
36. Tang, W., Sun, Y., Gu, Q., Li, Z.: Visual position prompt for mllm based visual grounding. IEEE Transactions on Multimedia (2026) 43
37. Team, G., et al.: Gemma 3 technical report (2025), https://arxiv.org/abs/2503.19786 9, 10, 11
38. Team, Q.: Qwen2.5-vl (January 2025), https://qwenlm.github.io/blog/qwen2.5-vl/ 2, 4, 9, 11
39. Team, Q.: Qwen3-vl technical report (2025), https://arxiv.org/abs/2511.21631v2 2, 4, 9, 10, 11
40. Tkachenko, M., Malyuk, M., Holmanyuk, A., Liubimov, N.: Label Studio: Data labeling software (2020-2025), https://github.com/HumanSignal/label-studio, open source software available from https://github.com/HumanSignal/label-studio 6
41. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., Lin, J.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024) 9, 10, 11
42. Wang, X., Wu, Z., Xie, J., Ding, Z., Yang, B., Li, Z., Liu, Z., Li, Q., Dong, X., Chen, Z., Wang, W., Zhao, X., Chen, J., Duan, H., Xie, T., Su, S., Yang, C., Yu, Y., Huang, Y., Liu, Y., Zhang, X., Yue, X., Su, W., Zhu, X., Shen, W., Dai, J., Wang, W.: Mmbench-gui: Hierarchical multi-platform evaluation framework for gui agents. arXiv preprint arXiv:2507.19478 (2025) 4
43. Wolf, I., et al.: The medical imaging interaction toolkit. Medical Image Analysis **9**(6), 594–604 (2005). https://doi.org/10.1016/j.media.2005.04.005 4, 7, 21
44. Wollny, G.: Ginkgo CADx: Open source medical imaging. https://ginkgo-cadx.com (2024) 4, 7, 21
45. Wu, Q., Cheng, K., Yang, R., Zhang, C., Yang, J., Jiang, H., Mu, J., Peng, B., Qiao, B., Tan, R., et al.: Gui-actor: Coordinate-free visual grounding for gui agents. Advances in Neural Information Processing Systems **38**, 15101–15128 (2026) 2, 4, 9, 10, 11, 30
46. Wu, Z., Wu, Z., Xu, F., Wang, Y., Sun, Q., Jia, C., Cheng, K., Ding, Z., Chen, L., Liang, P.P., et al.: Os-atlas: Foundation action model for generalist gui agents. In: International Conference on Learning Representations. vol. 2025, pp. 5090–5108 (2025) 9, 10, 11
47. Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., Xie, Z., Wu, Y., Hu, K., Wang, J., Sun, Y., Li, Y., Piao, Y., Guan, K., Liu, A., Xie, X., You, Y., Dong, K., Yu, X., Zhang, H., Zhao, L., Wang, Y., Ruan, C.: Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding (2024), https://arxiv.org/abs/2412.10302 9, 10, 11
48. Xie, T., Zhang, D., Chen, J., Li, X., Zhao, S., Cao, R., Hua, T.J., Cheng, Z., Shin, D., Lei, F., Liu, Y., Xu, Y., Zhou, S., Savarese, S., Xiong, C., Zhong, V., Yu, T.: Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In: Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Datasets and Benchmarks Track (2024), https://proceedings.neurips.cc/paper_files/paper/2024/hash/5d413e48f84dc61244b6be550f1cd8f5-Abstract-Datasets_and_Benchmarks_Track.html 2, 45
49. Xu, Y., Zhang, Y., Liu, J., Chen, J.: Structuring gui elements through vision language models: Towards action space generation (2025) 42

50. Xu, Y., Wang, Z., Wang, J., Lu, D., Xie, T., Saha, A., Sahoo, D., Yu, T., Xiong, C.: Aguvis: Unified pure vision agents for autonomous gui interaction. In: Proceedings of the 42nd International Conference on Machine Learning (ICML) (2025), to appear 9, 10, 11
51. Yue, J., Zhang, S., Jia, Z., Xu, H., Han, Z., Liu, X., Wang, G.: Medsg-bench: A benchmark for medical image sequences grounding. Advances in Neural Information Processing Systems **38** (2026) 3, 4
52. Yushkevich, P.A., Piven, J., Hazlett, H.C., Smith, R.G., Ho, S., Gee, J.C., Gerig, G.: User-guided 3d active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage **31**(3), 1116–1128 (2006) 4, 7, 21
53. Zhao, Y., Chen, W.N., Inan, H.A., Kessler, S., Wang, L., Wutschitz, L., Yang, F., Zhang, C., Minervini, P., Rajmohan, S., et al.: Learning gui grounding with spatial reasoning from visual feedback. arXiv preprint arXiv:2509.21552 (2025) 7
54. Zheng, B., Gou, B., Kil, J., Sun, H., Su, Y.: GPT-4V(ision) is a generalist web agent, if grounded. In: Proceedings of the 41st International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 235, pp. 61349–61385. PMLR (2024), https://proceedings.mlr.press/v235/zheng24e.html 2
55. Zhou, B., Huang, Z., Guo, Y., Gu, Z., Xia, T., Luo, Z., Tang, F., Kong, D., Shang, Y., Ou, S., Guo, Z., Meng, C., Shen, S.: Venusbench-gd: A comprehensive multi-platform gui benchmark for diverse grounding tasks (2025), https://arxiv.org/abs/2512.16501 4