EGOXR-GUI: BENCHMARKING GUI GROUNDING IN PHYSICAL–DIGITAL EXTENDED REALITY

Anonymous authors
Paper under double-blind review

ABSTRACT

The development of multimodal large language models has significantly advanced intelligent graphical user interface (GUI) agents. However, current research and benchmarking mainly focus on traditional platforms, like computer desktop and mobile interfaces, without considering the rapidly developing field of extended reality (XR). Unlike traditional platforms, XR requires models to understand the vital link between digital overlays and complex physical contexts. Moreover, unique platform artifacts and real-world egocentric interactions introduce additional challenges, such as perspective distortion, hand occlusion, and cluttered backgrounds. To this end, we present **EgoXR-GUI**, the first XR-specific GUI grounding benchmark, spanning explicit, spatial, and semantic grounding tasks, and featuring over 1,000 high-quality annotated samples collected across diverse real-world scenarios. Our results show that state-of-the-art models perform worse on these tasks than on conventional 2D grounding, revealing a critical gap in their ability to reason within hybrid physical-digital environments.

1 INTRODUCTION

The advances in Multimodal Large Language Models (MLLMs) (Bai et al., 2025b;a; Team et al., 2025; Yu et al., 2025; Chen et al., 2024c) have demonstrated remarkable capabilities in visual and language understanding, enabling significant progress in tasks such as visual question answering (Liu et al., 2023a;b; 2024), multimodal reasoning (Wang et al., 2025e; Zhang et al., 2023), and embodied decision-making (Black et al., 2026; Driess et al., 2023; Zitkovich et al., 2023). Meanwhile, graphical user interface (GUI) understanding with MLLMs has emerged as a prominent research direction, as grounding natural language instructions to visually localized interface elements enables a new generation of intelligent GUI agents capable of automated navigation and goal-oriented interaction in complex digital environments.

Recently, extended reality (XR) platforms have introduced a new form of user interface that seamlessly integrates the surrounding physical world with digital environments. However, despite substantial progress in GUI understanding across diverse platforms and operating systems, GUI grounding in XR platforms remains largely understudied. Compared to conventional GUI grounding tasks, XR GUIs are tightly integrated with the surrounding environment and pose fundamentally different challenges. Drawing on prior XR research, we synthesize four challenge dimensions for GUI grounding, as illustrated in Fig. 1:

- **Perceptual perturbation** caused by egocentric interaction and user surrounding scene context introduces occlusion and visual interference, complicating reliable perception of GUI elements (Macedo & Apolinario, 2023; Tsai et al., 2026).
- **Semantic ambiguity** arises from the co-presence of virtual interface elements and semantically similar scene entities in an egocentric context (Cao et al., 2026), and from context-dependent user intent (Kim et al., 2026a), complicating alignment between task-specific instructions and the correct interface components.
- **Spatial uncertainty** is induced by viewpoint-dependent interface placement in XR environments, whereas conventional GUI settings rely on spatial references fixed to stable 2D coordinate systems (Pei et al., 2024). Adaptive placement studies further connect layout structure to physical context (Zhou et al., 2026).

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

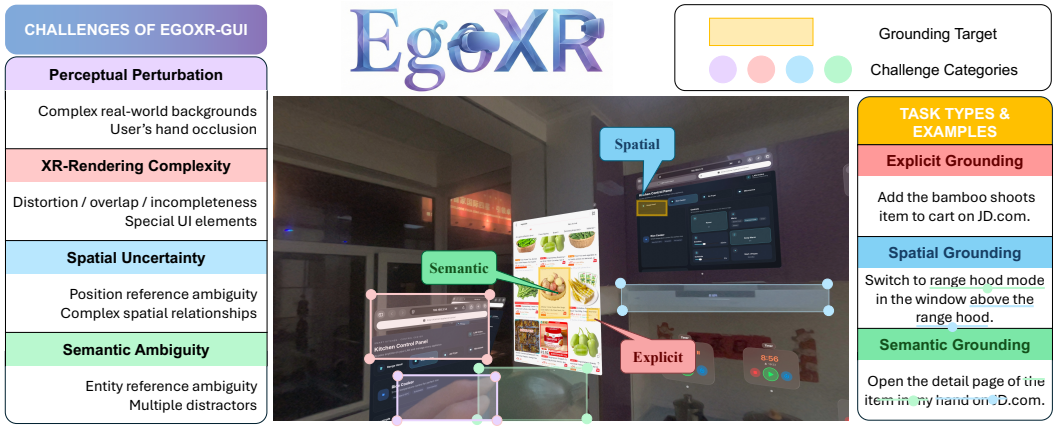


Figure 1: **Challenges and examples of our EgoXR-GUI benchmark.** Real-world XR interfaces introduce perceptual perturbations, rendering artifacts, spatial uncertainty, and semantic ambiguity due to the coupling between the physical world and digital content, making UI grounding substantially more challenging than conventional GUI understanding. Notably, EgoXR-GUI evaluates models across three task types—explicit, spatial, and semantic grounding—under these conditions.

- **XR-rendering complexity** stems from the unique interface characteristics, including overlapping or clustered windows, perspective distortion, partial visibility, and ultra-high-resolution rendering, introducing additional perceptual challenges for MLLMs (Zollmann et al., 2021).

These challenges motivate a dedicated benchmark to systematically evaluate existing GUI understanding approaches and assess whether current MLLMs can effectively reason about interfaces within hybrid digital–physical environments. To this end, we introduce EgoXR-GUI, the first benchmark for XR GUI grounding in egocentric daily contexts, where virtual interface elements are embedded within real-world interactions. Specifically, EgoXR-GUI covers three aspects: *Explicit*, *Spatial*, and *Semantic*. *Explicit* evaluates identifying interface elements explicitly referenced in instructions. *Spatial* assesses reasoning over relative positions between interface components and real-world entities during egocentric interaction. *Semantic* examines resolving ambiguous or abstract UI references from user-centric, task-driven prompts. Fig.1 visualizes these three task types.

We evaluate prevailing MLLMs and GUI grounding methods on EgoXR-GUI and provide a comprehensive analysis of their performance. Our results show that even state-of-the-art methods frequently fail on spatial and semantic grounding, due to the inherent challenges of interpreting hybrid physical–digital scenes and task-driven user queries. General proprietary models perform better on Semantic Grounding, likely benefiting from stronger alignment with real-world user queries. We further present detailed analyses and visualizations to highlight the key failure modes revealed by our benchmark. Finally, we explore two methods to enhance model performance, including (1) enhancing the agentic framework by introducing a task router to optimize the task processing workflow; (2) developing an automated data synthesis pipeline to address the scarcity of semantic grounding data, generating 500 synthetic XR images and 1,500 annotations. Fine-tuning Qwen3-VL-8B on this dataset yields a performance gain, with semantic grounding accuracy increasing from 23.17% to 41.46%. We demonstrate the practical utility of our benchmark by deploying it in a downstream XR GUI Agent application, shown in supplementary video 13:40 and Appendix C. The dataset and code are available for anonymous review.

2 Related Works

2.1 GUI Grounding

A GUI agent's operational workflow typically comprises planning, grounding, and action (Gonzalez-Pumariega et al., 2025; Li et al., 2026; Gu et al., 2025; Hong et al., 2024; Qin et al., 2025; Wang et al., 2025a; 2024; 2025f; Xu et al., 2024). Task execution accuracy and efficiency depend on UI grounding, which involves identifying interface elements from natural language instructions and

Table 1: Comparison of our Ego-GUI benchmark with existing GUI grounding and navigation benchmarks. We categorize them by environment, task types, and statistical scale.

Benchmarks	Environments	Tasks			Statistics		
	Platform	Explicit	Semantic	Spatial	#Sample	Avg.ele	Apps
ScreenSpot (Cheng et al., 2024a)	Desktop, Web, Mobile	✓	–	–	1,272	1.46	–
ScreenSpot V2 (Wu et al., 2024)	Desktop, Web, Mobile	✓	–	–	1,272	1.74	–
ScreenSpot-Pro (Li et al., 2025b)	Desktop	✓	–	–	1,581	1.0	23
OSWorld-G (Xie et al., 2025)	Desktop	✓	–	✓	564	2.25	8
UI-Vision (Nayak et al., 2025)	Desktop	✓	–	✓	5,479	4.64	83
VenusBench-GD (Zhou et al., 2025a)	Desktop, Web, Mobile	✓	–	✓	6,166	2.83	97
ScreenDrag (Liao et al., 2025)	Desktop	✓	–	✓	5,333	1.25	3
EgoXR-GUI (Ours)	XR	✓	✓	✓	1,068	4.54	47

GUI screenshots. Early grounding models utilize supervised fine-tuning on large-scale annotated screenshot datasets (Gou et al., 2025; Chen et al., 2026; Xie et al., 2025; Qin et al., 2025; Wu et al., 2025; Wang et al., 2025c). Inspired by DeepSeek-R1, recent approaches shift toward reinforcement learning frameworks using continuous or spatial reward signals to enhance robustness and precision (Tang et al., 2025b; Company, 2025a;b; Seed, 2025; Yang et al., 2025; Tang et al., 2025a; Lei et al., 2025; Zhou et al., 2025b). To address inaccurate localization and comprehension bias, several studies introduce training-free methods to further improve grounding accuracy (Zhang et al., 2025; 2026).

2.2 GUI Benchmarks

2.2.1 GUI Grounding and Agent Benchmarks

Existing GUI grounding benchmarks focus on traditional desktop and mobile environments. Early works (Cheng et al., 2024a; Wu et al., 2024) establish foundational evaluation methodology for element localization via screenshots and natural language instructions, while subsequent benchmarks increase complexity and scale: ScreenSpot-Pro (Li et al., 2025b) incorporates high-resolution professional software, WinSpot (Hui et al., 2025) targets Windows environments, OSWorld-G (Xie et al., 2025) structures grounding tasks from agent rollouts, and VenusBench-GD (Zhou et al., 2025a) and UI-Vision (Nayak et al., 2025) broaden platform coverage to web, mobile, and desktop. For GUI agent evaluation, benchmarks such as OSWorld (Xie et al., 2024), AndroidWorld (Rawles et al., 2024), and MMBench-GUI (Wang et al., 2025d) assess multi-step task completion pipelines spanning planning, grounding, and action across desktop and mobile platforms. Notably, prior studies (Xie et al., 2025; Wang et al., 2025d) confirm that grounding represents the fundamental bottleneck for GUI agents—a limitation we hypothesize extends to XR settings, motivating our dedicated grounding benchmark.

2.2.2 GUI Benchmarks including XR

Only a handful of works have examined GUI understanding in extended reality (XR) environments. GUI-World (Chen et al., 2024a) introduces a comprehensive, video-based multi-platform GUI benchmark that evaluates models through question answering tasks and partially incorporates XR settings. Orienter (Li et al., 2024a) establishes an early benchmark for interactive graphical element (IGE) detection within virtual reality (VR) applications. However, these studies primarily focus on XR interface understanding tasks rather than explicit grounding, as targeted in our benchmark. These GUI benchmarks do not systematically evaluate instruction-based localization of interface elements whose interpretation depends on the surrounding physical context during daily activities. To address this gap, we introduce a timely XR GUI grounding benchmark in egocentric contexts.

Prior XR research also studies how interfaces relate to their physical context. SituationAdapt adjusts layouts to environmental and social cues (Li et al., 2024b), while MineXR elicits personalized widgets and their placement in everyday environments (Cho et al., 2024). Satori uses actions, goals, and scene context for proactive assistance (Li et al., 2025a). ProCap separates physical scenes from projected content for captioning (Cao et al., 2026). SpeechLess resolves underspecified requests using personal context and spatial memory (Kim et al., 2026a). Zhou et al. (2026) model placement preferences

across physical environments, emphasizing layout structure and consistency. These studies connect interface use to situated activities and user intent, informing our grounding tasks.

2.3 EGOCENTRIC VQA BENCHMARKS

Recent works have explored egocentric benchmarks (Zhou et al., 2025c; Mangalam et al., 2023; Cheng et al., 2024b; Fan, 2019) to assess first-person visual understanding capabilities in multimodal large language models (MLLMs) (Shiri et al., 2024). EgoSchema (Mangalam et al., 2023) evaluates causal reasoning in egocentric narratives, while EgoPlan (Chen et al., 2024b) focuses on goal-oriented reasoning from ongoing activities. Peng et al. (2025) incorporate gaze signals to assess intention-aware reasoning. EgoDynamics4D (Huang et al., 2025) extends egocentric understanding to 3D spatial grounding of objects and agents. Although these benchmarks advance human-centric perception and embodied reasoning, they primarily focus on activity or object-level spatial reasoning. Our work evaluates whether MLLMs can localize GUI elements in egocentric contexts, disambiguating them from surrounding objects under viewpoint-dependent spatial and semantic ambiguity.

3 EGOXR-GUI

We denote an egocentric observation in an XR environment as I. Notably, I is composed of both physical entities $\mathcal{O} = \{o_1, \ldots, o_n\}$ and virtual UI elements $\mathcal{U} = \{u_1, \ldots, u_m\}$. Given I and a natural language query q conditioned on the user's current activity, the objective is to identify the target UI element $u^* \in \mathcal{U}$ and predict its corresponding 2D bounding box on the image plane of I.

3.1 BENCHMARK DESIGN

The design of EgoXR-GUI tasks originates from authentic XR usage scenarios. First-person XR devices are distinguished by the integration of virtual UI windows with real-world backgrounds. Consequently, XR interface elements are often closely coupled with real-world contextual information. Leveraging this characteristic, we define three categories of UI element grounding tasks in EgoXR-GUI, classified by the relationship between the input instruction and the real-world background information. All three are single-step localization tasks; they differ in the evidence needed to identify the target. This distinction is consistent with XR studies of attribute-based and relational references, and context-dependent user intentions (Kim et al., 2026b; Li et al., 2025a).

• Explicit Grounding: Input user query q specifies identifiable attributes of the target UI element $u^* \in \mathcal{U}$ (e.g., text labels or explicit UI descriptions). Similar to prior GUI grounding benchmarks, this task setting evaluates whether models can accurately localize UI elements despite the unique rendering, real-world background interference, and perceptual complexity of XR environments.

• Semantic Grounding: Instructions specify a UI window but lack the complete semantic information to determine u^* from $\mathcal{U}$ (*e.g.*, "open the detail page for the item in my hand in xx shopping software"), requiring the model to extract semantic cues from the real-world background to select the correct UI element. This setting evaluates the model's ability to perform joint reasoning over the visual context in I and the implicit query q to resolve the correct target.

• Spatial Grounding: The query q specifies the target UI element $u^* \in \mathcal{U}$ through a spatial relation with respect to a reference entity $o_j \in \mathcal{O}$ (e.g., "above the xx" or "on the wall to the right"). Resolving u^* therefore requires reasoning over relative spatial configuration in the egocentric observation I, including depth, surface orientation, and reference object identification, distinct from previous benchmarks (Nayak et al., 2025; Zhou et al., 2025a; Xie et al., 2025; Liao et al., 2025) that are largely confined to 2D planar spatial reasoning.

Fig. 1 illustrates examples and challenges of the three task categories.

3.2 BENCHMARK CHARACTERISTICS

Prioritizing the practical utility of GUI agents, we design a structured benchmark. Based on recent egocentric video dataset capture protocol (Grauman et al., 2022) and typical XR use cases, we select six representative scenarios subdivided into ten specific places. We annotate tasks from the most representative activities at each location, encompassing 20+ activities across 47 XR applications and

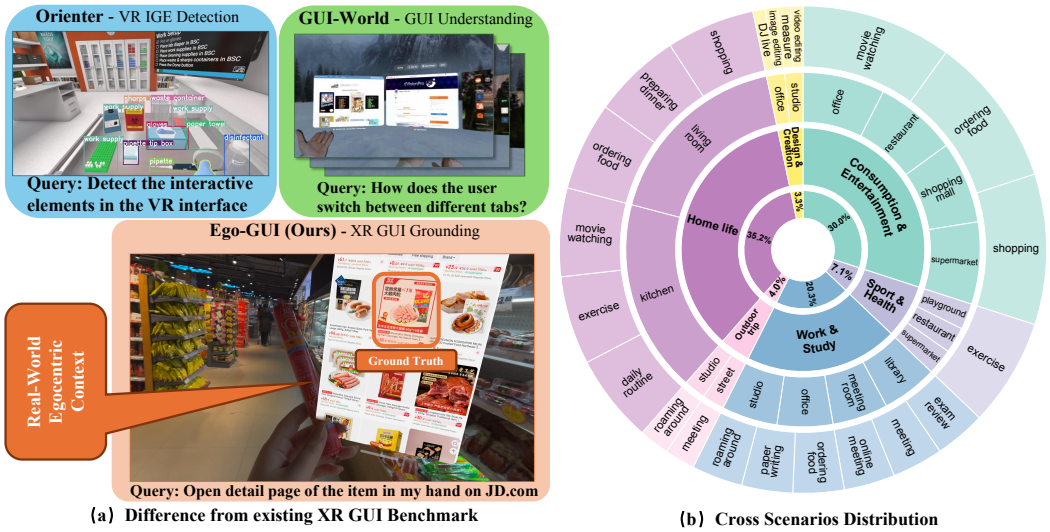


Figure 2: **Overview of the proposed benchmark**. (*a*) Comparison with existing XR GUI benchmarks: Unlike **Orienter** (Li et al., 2024a) which focuses on virtual IGE detection and **GUI-World** (Chen et al., 2024a) which targets general GUI understanding, **EgoXR-GUI** uniquely emphasizes XR GUI grounding within real-world egocentric contexts. (*b*) The hierarchical distribution of our dataset across diverse scenarios (e.g., Consumption, Home life), and daily activities, illustrates the comprehensiveness of the benchmark.

interfaces. This design supports evaluation in authentic environments. Everyday AR studies and personalized interface mining provide related examples of media, productivity, wellness, and context-dependent interface use (Tran et al., 2025; Cho et al., 2024; Lu & Bowman, 2021). Appendix A.4 summarizes the literature connections and functional interface coverage.

As shown in Table 1, we compare EgoXR-GUI with prominent existing GUI grounding benchmarks across various platforms. EgoXR-GUI comprises 1068 annotated task samples, maintaining a data scale comparable to prior benchmarks. Furthermore, the unique interface of XR introduces diverse perspectives, backgrounds, and lighting conditions, which further enrich the dataset. We also compare our data samples with those from previous XR-related GUI benchmarks in Fig. 2 (a).

3.3 Dataset Building Pipeline

3.3.1 Data Collection

We selected Apple Vision Pro as our data collection platform due to its sophisticated UI, diverse applications, and high-fidelity reality passthrough. To obtain clear screenshots without foveated rendering, we utilized official developer tools (Apple Inc., 2024). Users are first asked to select locations and activities where they would naturally use XR devices for entertainment or productivity. During each session, users interact with XR interfaces from an egocentric perspective while performing the selected activities. For each scenario, participants are instructed to select a UI window or interface element with a specific task goal in mind, mimicking real-world XR usage. This goal-driven interaction setup ensures that samples reflect natural user intent and realistic XR application scenarios.

Since certain data were collected in public spaces, we mask all privacy-sensitive information within the images to protect privacy, ensuring that such processing does not impede model prediction performance. We recruited eight volunteers to wear the device for data collection, capturing varied perspectives, fixation habits, and hand appearances.

3.3.2 Data Annotation

The subjects initially recorded their task intentions based on their authentic task goals. Two researchers, both members of the author team, then standardized the instruction wording and manually

annotated all data via Label Studio (Tkachenko et al., 2020-2025) to ensure quality and semantic standardization. We maintained high annotation accuracy by defining unified templates that are concise yet uniquely mapped to the ground truth. To ensure diversity, we annotated 3 to 5 tasks per image, with most samples covering at least two task categories. To ensure annotation accuracy and consistency, the two researchers labeled the data and conducted mutual cross-reviews within the platform. In an additional audit, each researcher reviewed 100 annotations from the other, and one original participant checked 100 samples from their own data. Criterion-level pass rates ranged from 99% to 100%; the criteria and subset results are reported in Table 5.

4 Experiments

4.1 Experiment Settings

Our evaluation covers three categories of baselines: closed-source MLLMs, open-source pretrained VLMs at various scales, and open-source models specifically designed for GUI grounding that achieve leading results on benchmarks such as ScreenSpot Pro (Li et al., 2025b), ScreenSpot-v2 (Wu et al., 2024), and UI-Vision (Nayak et al., 2025). All models are evaluated using their official prompt templates; details are provided in the Appendix. Each screenshot evaluates grounding at a single decision point. Following ScreenSpot (Cheng et al., 2024a) and ScreenSpot Pro (Li et al., 2025b), a prediction is correct if the predicted coordinates fall within the ground truth bounding box. We report accuracy separately for text and icon elements to enable fine-grained analysis. We also evaluate human performance for comparison(specific details are provided in E.2).

4.2 Benchmark Results

Experimental results are summarized in Tab. 2. General-purpose multimodal large language models (MLLMs) demonstrate competitive performance on simple GUI grounding tasks. Notably, recent models like Qwen3-VL-8B (Bai et al., 2025a) achieve strong results in Explicit Grounding, reaching 84.04 average accuracy and matching several specialized GUI expert models.

Furthermore, while the gap between RL-based and SFT-based methods is modest on existing benchmarks, RL-based approaches perform better on our benchmark. We hypothesize that this reflects differing optimization dynamics: SFT-based models may overfit training distributions, whereas RL-based training promotes broader exploration and robust policy learning, enabling better generalization to diverse EgoXR-GUI scenarios.

Across the three task types, Explicit Grounding is noticeably easier than Spatial Grounding and Semantic Grounding, as the latter two require additional reasoning beyond direct element identification. We also observe modality-dependent differences: text elements are generally easier for explicit and

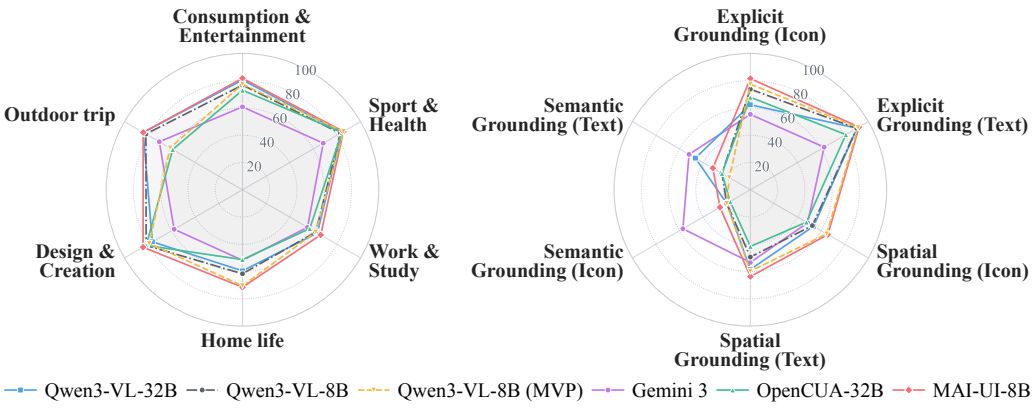


Figure 3: Accuracy (%) of six representative models by scenario (*left*) and task type (*right*). Full results are in Table 2.

Table 2: Experimental results on different task types. We evaluate 28 models across five categories on three grounding tasks. We also evaluate human performance for comparison. All numbers are reported in percentage (%).

Models	Explicit Grounding			Spatial Grounding			Semantic Grounding			Overall		
	text	icon	avg	text	icon	avg	text	icon	avg	text	icon	avg
Closed-source Models												
Gemini 3 (Google, 2026)	62.50	55.35	59.83	**53.25**	48.48	**50.57**	**51.94**	**57.14**	**53.05**	59.43	53.83	57.30
Seed1.6 VL (Seed, 2025)	**74.56**	**66.05**	**71.39**	33.77	**54.55**	45.45	**51.94**	40.00	49.39	**65.46**	**60.99**	**63.76**
General Open-source VLMs												
Qwen2.5-VL-7B-Instruct (Bai et al., 2025b)	71.93	49.45	63.55	28.57	32.32	30.68	25.58	2.86	20.73	57.77	41.23	51.50
Qwen3-VL-2B (Bai et al., 2025a)	80.70	55.72	71.39	27.27	36.36	32.39	15.50	11.43	14.63	61.69	47.16	56.18
Qwen3-VL-8B (Bai et al., 2025a)	90.13	**73.80**	**84.04**	49.35	52.53	51.14	24.03	20.00	23.17	72.55	63.95	69.29
Qwen3-VL-32B (Bai et al., 2025a)	**90.35**	62.36	79.92	**58.44**	**53.54**	**55.68**	**46.51**	20.00	**40.85**	**78.13**	56.54	**69.94**
Qwen3-VL-30B-A3B (Bai et al., 2025a)	86.84	71.96	81.29	46.75	43.43	44.89	20.93	17.14	20.12	69.38	60.25	65.92
InternVL3.5-8B (Wang et al., 2025b)	60.53	36.53	51.58	22.08	32.32	27.84	16.28	11.43	15.24	47.36	33.33	42.04
InternVL3.5-30B-A3B (Wang et al., 2025b)	63.82	44.28	56.53	27.27	30.30	28.98	20.16	11.43	18.29	50.98	38.02	46.07
MiniCPM-V4.5-8B (Yu et al., 2025)	50.00	28.41	41.95	11.69	13.13	12.50	13.95	2.86	11.59	38.46	22.47	32.40
GUI-specific Models (SFT)												
OS-Atlas-7B (Wu et al., 2024)	51.32	34.69	45.12	16.88	21.21	19.32	11.63	2.86	9.76	39.52	28.64	35.39
UGround-7B (Gou et al., 2025)	37.06	23.99	32.19	11.69	14.14	13.07	17.05	2.86	14.02	30.17	19.75	26.22
CogAgent-9B (Hong et al., 2024)	30.04	23.99	27.79	6.49	17.17	12.50	7.75	0.00	6.10	23.08	20.25	22.00
OpenCUA-7B (Wang et al., 2025c)	73.25	54.61	66.30	33.77	43.43	39.20	11.63	8.57	10.98	56.56	47.90	53.28
OpenCUA-32B (Wang et al., 2025c)	**80.92**	**67.90**	**76.07**	**41.56**	47.47	**44.89**	**24.03**	**17.14**	**22.56**	**65.16**	**58.52**	**62.64**
OpenCUA-72B (Wang et al., 2025c)	55.26	56.09	55.57	37.66	**51.52**	**44.89**	19.38	5.71	16.46	46.30	50.62	47.94
GUI-specific Models (RL)												
UI-TARS 1.5-7B (Seed, 2025)	69.96	51.66	63.14	19.48	26.26	23.30	10.85	11.43	10.98	52.49	41.98	48.50
Tianxi-Action-7B (Tang et al., 2025b)	79.39	56.83	70.98	35.06	30.30	32.39	21.71	11.43	19.51	63.05	46.42	56.74
Holo1.5-7B (Company, 2025a)	86.84	71.96	81.29	36.36	54.55	46.59	18.60	14.29	17.68	67.57	62.72	65.73
Holo1.5-72B (Company, 2025a)	88.15	80.44	85.28	53.25	68.69	61.93	**43.41**	20.00	**38.41**	75.41	72.35	74.25
Holo2-4B (Company, 2025b)	78.07	61.25	71.80	29.87	39.39	35.23	14.73	11.43	14.02	60.18	51.60	56.93
Holo2-8B (Company, 2025b)	80.48	64.94	74.69	38.96	52.53	46.59	20.16	14.29	18.90	63.80	57.53	61.42
Holo2-30B-A3B (Company, 2025b)	89.91	81.55	86.80	58.44	67.68	63.64	24.81	20.00	23.78	73.60	72.84	73.31
UI-Venus 1.5-8B (Team et al., 2026)	**92.11**	83.76	88.03	**63.64**	74.74	**69.89**	31.01	20.00	28.66	75.87	76.05	**75.94**
UI-Venus 1.5-30B-A3B (Team et al., 2026)	89.69	**84.13**	87.62	55.84	**76.77**	67.61	25.58	17.14	23.78	73.30	**76.54**	74.53
MAI-UI-8B (Zhou et al., 2025b)	**92.11**	81.55	**88.17**	**63.64**	65.66	64.77	31.78	**25.71**	30.49	**77.07**	72.84	75.47
Training-Free Method												
Qwen3-VL-8B (MVP) (Bai et al., 2025a)	**91.23**	**77.49**	**86.11**	**59.74**	**64.65**	**62.50**	17.83	**20.00**	18.29	**73.30**	**69.38**	**71.82**
Tianxi-Action-7 (BAMI) (Tang et al., 2025b)	80.26	59.04	72.35	51.95	40.40	45.45	**24.81**	11.43	**21.95**	66.21	50.37	60.21
Human Performance	99.56	100.00	99.72	100.00	96.97	98.30	99.23	100.00	99.39	99.40	99.51	99.44

semantic grounding, whereas icons tend to be easier for spatial grounding. Textual labels likely facilitate direct identification and instruction interpretation through explicit semantic cues.

We show the per-scenario performance breakdown in Fig. 3 (left) and observe a consistent performance drop on Home-life and Work&Study across most models. We speculate that this is because the physical environments in these scenarios are inherently more complex and often contain multiple entities that are difficult to disambiguate semantically or spatially from the given prompt. From the overall performance in Fig. 3 (right) and Tab. 2, we observe that Gemini achieves higher performance on semantic grounding tasks. This suggests that proprietary models can more effectively reason about ambiguous references, likely due to stronger multimodal pretraining and access to larger-scale user preference data, highlighting a key bottleneck for open-source models in resolving such references. Results for Chinese instructions and efficiency are in Appendix B.5. We examine selected Semantic Grounding failures across headsets and the use of gaze coordinates in Appendices B.7 and B.8.

4.3 Analysis

We analyze MAI-UI 8B errors using cross-modal self-attention maps. Error case analysis of explicit and spatial grounding can be found in B.1 and B.2. Table 10 connects these examples and the semantic errors below to the four XR challenge dimensions.

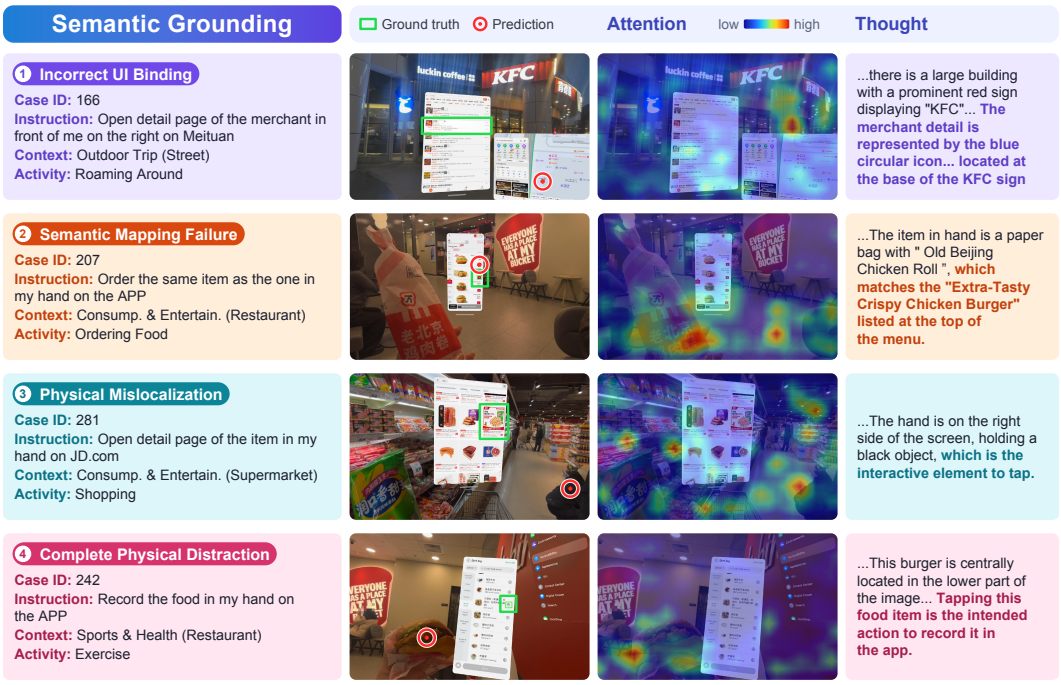


Figure 4: **Error case analysis of semantic grounding.** MAI-UI 8B exhibits *Incorrect UI Binding* (Case 166), *Semantic Mapping Failure* (Case 207), *Physical Mislocalization* (Case 281), and *Complete Physical Distraction* (Case 242). Each row shows the instruction, predicted target, attention map, and reasoning excerpt; green boxes mark ground truth and red circles mark predictions.

Error analysis on semantic grounding. Semantic Grounding requires identifying the referenced real-world object's semantic information and grounding the matching UI element within the target window. Accordingly, we categorize Semantic Grounding errors into four types:

• *Incorrect UI Binding*: In multi-window scenarios, the model correctly reasons about real-world semantics but selects the wrong UI window (Fig. 4, Case 166). Like Explicit Grounding errors, this indicates UI comprehension ambiguity. Processing real-world semantics distracts the model from UI windows, amplifying this weakness.

• *Semantic Mapping Failure*: The model's response deviates from the real-world object's actual semantics (Fig. 4, Case 207). Attention heatmaps show focus on both the object and target window, yet biased understanding leads the model to ground a misleading element.

• *Physical Mislocalization*: The model fails to identify the correct real-world object, as shown by the reasoning descriptions and attention heatmaps in Fig. 4 (Case 281).

• *Physical Distraction*: The model fixates on physical objects and fails to recognize any UI windows (Fig. 4, Case 242).

The latter three types stem from deficient real-world semantic comprehension. Combined with tabular results, we attribute this to GUI-specific base models and their training paradigms. The scarcity of XR GUI training data prevents these models from simultaneously improving GUI grounding and foundational physical-world understanding.

4.4 Agentic Inference Framework

Our analysis (detailed in Appendix B.4) shows that Qwen Thinking versions improve average Semantic Grounding accuracy at all four evaluated scales, motivating the use of a reasoning pathway for queries that require physical-scene understanding.

Table 3: Performance comparison of baseline models and different inference frameworks on EgoXR-GUI. All numbers are reported in percentage (%).

Models	Explicit Grounding			Spatial Grounding			Semantic Grounding			Overall		
	text	icon	avg	text	icon	avg	text	icon	avg	text	icon	avg
Baseline Models												
Seed1.6 VL	74.56	66.05	71.39	33.77	54.55	45.45	**51.94**	**40.00**	**49.39**	65.46	60.99	63.76
MAI-UI-8B	**92.11**	**81.55**	**88.17**	**63.64**	**65.66**	**64.77**	31.78	25.71	30.49	**77.07**	**72.84**	**75.47**
Framework-based Methods												
Color BBox Mark(Seed1.6 VL + MAI-UI-8B)	81.57	75.27	79.22	51.94	58.58	55.68	58.13	37.14	53.65	73.60	67.90	71.44
AgentS3 (Gonzalez-Pumariega et al., 2025) (Seed1.6 VL + MAI-UI-8B)	84.42	**69.74**	78.95	51.94	65.65	59.65	64.34	37.14	**58.53**	76.77	65.92	72.65
TaskRouter (Seed1.6 VL + MAI-UI-8B)	**87.50**	**75.65**	**83.08**	**59.74**	**72.73**	**67.05**	55.81	**48.57**	54.27	**77.98**	72.59	**75.93**

Table 4: Synthetic-data fine-tuning of Qwen3-VL-8B-Instruct, evaluated on the real EgoXR-GUI Semantic Grounding set. Accuracy is reported in %; changes are percentage points.

Setting	Text	Icon	Overall
Base model	24.03	20.00	23.17
Fine-tuned on synthetic data	48.06	17.14	41.46
Change	+24.03	−2.86	+18.29

However, as shown in Tab. 3, the Color BBox Mark and AgentS3 frameworks improve Semantic Grounding over MAI-UI-8B while reducing its Explicit and Spatial Grounding accuracy. This trade-off motivates selecting the reasoning pathway according to the task.

To address this dilemma, we propose a training-free **Agentic Inference Framework** with a task-oriented routing mechanism. A VLM-based router dynamically activates a complex reasoning pathway exclusively for Semantic Grounding queries. For Explicit and Spatial tasks, it defaults to direct prediction. This zero-shot strategy selectively applies semantic reasoning and achieves the highest average accuracy among the frameworks compared in Table 3. For specific details, see B.3.

4.5 Data Synthesis and Fine-tuning For Semantic Grounding

The high cost of collecting real-world egocentric XR data prevents resolving grounding capability deficits through standard training. To overcome this constraint, we construct an automated data synthesis and annotation pipeline. Building on the EPIC-KITCHENS-100 video dataset (Damen et al., 2022), we employ generative AI (OpenAI, 2026) and multimodal large language models (Qwen Team, 2026) to render synthetic XR interface overlays onto physical reality frames. We then prompt the multimodal models to generate paired explicit and semantic grounding instructions, resulting in 500 synthetic XR images and 1,500 semantic grounding annotations. For bounding-box labels, we use UI-Venus-1.5-8B (Team et al., 2026), which achieves 92.11% Explicit Text accuracy and 88.03% Explicit average accuracy on our benchmark, with the corresponding explicit instructions. Two human reviewers audited 300 synthetic annotations; instruction validity and bounding-box accuracy both reached 96% (Appendix E.3). We subsequently fine-tune Qwen3-VL-8B-Instruct (Bai et al., 2025a) on this synthesized dataset and evaluate it on the real EgoXR-GUI Semantic Grounding set using the same metric and text/icon categories as the baseline. Experimental results show that fine-tuning increases semantic grounding accuracy from 23.17% to 41.46%. As shown in Table 4, the gain comes from text targets, while icon accuracy decreases from 20.00% to 17.14%.

5 Conclusion

In this paper, we present EgoXR-GUI, the first benchmark for GUI grounding in hybrid physical–digital extended reality (XR) environments. Our results show that while current models perform reasonably well on explicit grounding, they struggle with spatial and semantic grounding, where resolving relative geometry and ambiguous user references in egocentric views becomes essential. We also demonstrate that fine-tuning on synthetic XR data can improve average semantic grounding accuracy. We believe our work represents an essential step toward personal AI agents on XR devices that can interpret user intentions under inherent uncertainty.

AI USE STATEMENT

We used generative AI to construct the synthetic training data described in Section 4.5 and Appendix E.3. Qwen3.6-Plus performed scene analysis and generated candidate applications, image-generation prompts, and paired explicit and semantic instructions; GPT Image 2 generated XR interface overlays on EPIC-KITCHENS-100 frames; and UI-Venus-1.5-8B predicted target bounding boxes from the corresponding explicit instructions. This pipeline produced 500 synthetic XR images and 1,500 semantic grounding annotations. The models evaluated as baselines and used in the agentic inference frameworks are described in Section 4 and Appendix B.3.

For the original study, the researchers developed the research motivation, designed the benchmark, experiments, and synthesis pipeline, collected and annotated the real benchmark data, and interpreted the results without generative AI assistance. Mathematical theorem formulation and proof writing are not applicable to this empirical study. We additionally used AI for literature discovery, software coding assistance, and manuscript preparation, including LaTeX conversion, citation formatting, drafting these statements, and integrating the documented experiments and clarifications into the manuscript. Methodological decisions and final reference selection remained the responsibility of the authors.

Human verification included an audit of 300 synthetic annotations by two reviewers. They checked whether each instruction was supported by the physical scene and interface and identified a valid semantic target, and whether the predicted bounding box covered the intended GUI element. Instruction validity and bounding-box accuracy were each 96% (288/300) on this audited subset. The authors are responsible for the final manuscript, its scientific claims, and the data and code artifacts, including content produced with AI assistance.

ETHICS STATEMENT

EgoXR-GUI uses egocentric XR recordings that can expose personal environments, interface content, and bystanders. Eight volunteers contributed data after providing informed consent about the recording scope. Three participants who were not involved in annotation provided informed consent and performed the human evaluation described in Appendix E.2. Privacy-sensitive information was masked, including identifiable faces in public scenes, and the researchers responsible for annotation checked the processed data. Access to raw recordings was restricted to the collectors and researchers responsible for processing sensitive information; only processed data are intended for release (Appendix A). Data and artifact releases are limited to what participant authorization and the applicable licenses permit. In particular, the EPIC-KITCHENS-100 material used for synthesis retains its non-commercial license restrictions; these restrictions are not replaced by the release terms for the separately collected benchmark.

The benchmark can support more accessible and intuitive XR interfaces, but egocentric agents can also compromise privacy or cause distraction and unintended actions when grounding fails. Our experiments assess grounding accuracy and do not establish deployment safety. The small participant pool, collection of the main benchmark on a single headset, and static-frame design also limit generalization across users, devices, and interaction settings. These limitations and potential impacts are discussed in Appendices F and G; deployment requires further safety and privacy evaluation.

REPRODUCIBILITY STATEMENT

Section 3.3 and Appendix A describe data collection, privacy processing, and annotation; Section 4.1 defines the evaluation protocol and accuracy metric. Appendix E provides model prompts, the human-evaluation procedure, and the data-synthesis and fine-tuning configurations, including LoRA, image preprocessing, and computational resources. Appendix B.3 describes the agentic inference methods, and Appendices B.7 and B.8 describe the cross-device and gaze studies. The benchmark data[1] and code for evaluation, synthesis, and fine-tuning[2] are available through anonymous repositories. For the fine-tuning study, the synthetic images and annotations are used only for training; evaluation uses the real EgoXR-GUI Semantic Grounding set with the same metric and text/icon categories as the

[1] https://huggingface.co/datasets/Anonymous114/EgoXR-GUI

[2] https://anonymous.4open.science/r/EgoXR-GUI-8F69/

baseline, and includes no synthetic samples. Reproducing experiments that use proprietary models depends on continued API access and may be affected by provider-side updates.

REFERENCES

Apple Inc. *Capturing screenshots and video from your Apple Vision Pro for 2D viewing*. Apple Developer Documentation, 2024. URL https://developer.apple.com/documentation/visionos/capturing-screenshots-and-video-from-your-apple-vision-pro-for-2d-viewing. Accessed: 2024-05-20.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control, 2026. URL https://arxiv.org/abs/2410.24164.

Zimo Cao, Yuchen Deng, Haibin Ling, and Bingyao Huang. ProCap: Projection-Aware Captioning for Spatial Augmented Reality. In *2026 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 22–32. IEEE, March 2026. doi: 10.1109/vr67842.2026.00026. URL http://dx.doi.org/10.1109/VR67842.2026.00026.

Dongping Chen, Yue Huang, Siyuan Wu, Jingyu Tang, Liuyi Chen, Yilin Bai, Zhigang He, Chenlong Wang, Huichi Zhou, Yiqiang Li, Tianshuo Zhou, Yue Yu, Chujie Gao, Qihui Zhang, Yi Gui, Zhen Li, Yao Wan, Pan Zhou, Jianfeng Gao, and Lichao Sun. Gui-world: A dataset for gui-oriented multimodal llm-based agents, 2024a.

Jikai Chen, Long Chen, Dong Wang, Qinglin Su, Zhixuan Chu, Bingguang Hao, Leilei Gan, Chenyi Zhuang, and Jinjie Gu. V2p: Visual attention calibration for gui grounding via background suppression and center peaking, 2026. URL https://arxiv.org/abs/2508.13634.

Yi Chen, Yuying Ge, Yixiao Ge, Mingyu Ding, Bohao Li, Rui Wang, Ruifeng Xu, Ying Shan, and Xihui Liu. Egoplan-bench: Benchmarking multimodal large language models for human-level planning, 2024b. URL https://arxiv.org/abs/2312.06722.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024c.

Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Yantao Li, Jianbing Zhang, and Zhiyong Wu. Seeclick: Harnessing gui grounding for advanced visual gui agents, 2024a.

Sijie Cheng, Zhicheng Guo, Jingwen Wu, Kechen Fang, Peng Li, Huaping Liu, and Yang Liu. Egothink: Evaluating first-person perspective thinking capability of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14291–14302, June 2024b.

Hyunsung Cho, Yukang Yan, Kashyap Todi, Mark Parent, Missie Smith, Tanya R. Jonker, Hrvoje Benko, and David Lindlbauer. MineXR: Mining Personalized Extended Reality Interfaces. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, pp. 1–17. ACM, May 2024. doi: 10.1145/3613904.3642394. URL http://dx.doi.org/10.1145/3613904.3642394.

H Company. Holo1.5 - open foundation models for computer use agents, 2025a. URL https://huggingface.co/collections/Hcompany/holo15-68c1a5736e8583a309d23d9b.

H Company. Holo2 - open foundation models for navigation and computer use agents, 2025b. URL https://huggingface.co/collections/Hcompany/holo2.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130:33–55, 2022. URL https://doi.org/10.1007/s11263-021-01531-2.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

Chenyou Fan. Egovqa - an egocentric video question answering benchmark dataset. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 4359–4366, 2019. doi: 10.1109/ICCVW.2019.00536.

Gonzalo Gonzalez-Pumariega, Vincent Tu, Chih-Lun Lee, Jiachen Yang, Ang Li, and Xin Eric Wang. The unreasonable effectiveness of scaling agents for computer use, 2025. URL https://arxiv.org/abs/2510.02250.

Google. Gemini 3: Large language model, 2026. URL https://gemini.google.com/. Computer software. Version updated March 2026.

Boyu Gou, Ruohan Wang, Boyuan Zheng, Yanan Xie, Cheng Chang, Yiheng Shu, Huan Sun, and Yu Su. Navigating the digital world as humans do: Universal visual grounding for GUI agents. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=kxnoqaisCT.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

Zhangxuan Gu, Zhengwen Zeng, Zhenyu Xu, Xingran Zhou, Shuheng Shen, Yunfei Liu, Beitong Zhou, Changhua Meng, Tianyu Xia, Weizhi Chen, et al. Ui-venus technical report: Building high-performance ui agents with rft. *arXiv preprint arXiv:2508.10833*, 2025.

Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, et al. Cogagent: A visual language model for gui agents. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14281–14290, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Junsheng Huang, Shengyu Hao, Bocheng Hu, and Gaoang Wang. Understanding dynamic scenes in ego centric 4d point clouds. *ArXiv*, abs/2508.07251, 2025. URL https://api.semanticscholar.org/CorpusID:280567307.

Zheng Hui, Yinheng Li, Dan Zhao, Colby Banbury, Tianyi Chen, and Kazuhito Koishida. Winspot: Gui grounding benchmark with multimodal large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1086–1096, 2025.

Yoonsang Kim, Devshree Jadeja, Divyansh Pradhan, Yalong Yang, and Arie E. Kaufman. SpeechLess: Micro-utterance with Personalized Spatial Memory-aware Assistant in Everyday Augmented Reality. In *2026 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 217–227. IEEE, March 2026a. doi: 10.1109/vr67842.2026.00044. URL http://dx.doi.org/10.1109/VR67842.2026.00044.

Yoonsang Kim, Divyansh Pradhan, Devshree Jadeja, and Arie E. Kaufman. From Speech-to-Spatial: Grounding Utterances on A Live Shared View with Augmented Reality. In *2026 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 228–238. IEEE, March 2026b. doi: 10.1109/vr67842.2026.00045. URL http://dx.doi.org/10.1109/VR67842.2026.00045.

Geonsun Lee, Min Xia, Nels Numan, Xun Qian, David Li, Yanhe Chen, Achin Kulshrestha, Ishan Chatterjee, Yinda Zhang, Dinesh Manocha, David Kim, and Ruofei Du. Sensible Agent: A Framework for Unobtrusive Interaction with Proactive AR Agents. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*, UIST '25, pp. 1–22. ACM, September 2025. doi: 10.1145/3746059.3747748. URL http://dx.doi.org/10.1145/3746059.3747748.

Bin Lei, Nuo Xu, Ali Payani, Mingyi Hong, Chunhua Liao, Yu Cao, and Caiwen Ding. GUI-Spotlight: Adaptive Iterative Focus Refinement for Enhanced GUI Visual Grounding. *arXiv preprint arXiv:2510.04039*, 2025.

Chenyi Li, Guande Wu, Gromit Yeuk-Yin Chan, Dishita G Turakhia, Sonia Castelo Quispe, Dong Li, Leslie Welch, Claudio Silva, and Jing Qian. Satori: Towards proactive ar assistant with belief-desire-intention user modeling, 2025a. URL https://arxiv.org/abs/2410.16668.

Kaixin Li, Ziyang Meng, Hongzhan Lin, Ziyang Luo, Yuchen Tian, Jing Ma, Zhiyong Huang, and Tat-Seng Chua. Screenspot-pro: Gui grounding for professional high-resolution computer use. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 8778–8786, 2025b.

Shuqing Li, Binchang Li, Yepang Liu, Cuiyun Gao, Jianping Zhang, Shing-Chi Cheung, and Michael R Lyu. Grounded gui understanding for vision-based spatial intelligent agent: Exemplified by extended reality apps. *arXiv preprint arXiv:2409.10811*, 2024a.

Sijia Li, Xiaoyu Tan, Shahir Ali, Niels Schmidt, Gengchen Ma, and Xihe Qiu. Curiosity driven knowledge retrieval for mobile agents. *arXiv preprint arXiv:2601.19306*, 2026.

Zhipeng Li, Christoph Gebhardt, Yves Inglin, Nicolas Steck, Paul Streli, and Christian Holz. SituationAdapt: Contextual UI Optimization in Mixed Reality with Situation Awareness via LLM Reasoning. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, UIST '24, pp. 1–13. ACM, October 2024b. doi: 10.1145/3654777.3676470. URL http://dx.doi.org/10.1145/3654777.3676470.

Zeyi Liao, Yadong Lu, Boyu Gou, Huan Sun, and Ahmed Awadallah. Beyond clicking: A step towards generalist gui grounding via text dragging. *arXiv preprint arXiv:2601.06031*, 2025.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023b.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Feiyu Lu and Doug A. Bowman. Evaluating the Potential of Glanceable AR Interfaces for Authentic Everyday Uses. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pp. 768–777. IEEE, March 2021. doi: 10.1109/vr50410.2021.00104. URL http://dx.doi.org/10.1109/VR50410.2021.00104.

Marcio C. F. Macedo and Antonio L. Apolinario. Occlusion Handling in Augmented Reality: Past, Present and Future. *IEEE Transactions on Visualization and Computer Graphics*, 29(2): 1590–1609, February 2023. ISSN 2160-9306. doi: 10.1109/tvcg.2021.3117866. URL http://dx.doi.org/10.1109/TVCG.2021.3117866.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *ArXiv*, abs/2308.09126, 2023. URL https://api.semanticscholar.org/CorpusID:261031047.

Mykola Maslych, Difeng Yu, Amirpouya Ghasemaghaei, Yahya Hmaiti, Esteban Segarra Martinez, Dominic Simon, Eugene M. Taranta, Joanna Bergström, and Joseph J. LaViola. From Research to Practice: Survey and Taxonomy of Object Selection in Consumer VR Applications. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 990–999. IEEE, October 2024. doi: 10.1109/ismar62088.2024.00115. URL http://dx.doi.org/10.1109/ISMAR62088.2024.00115.

Shravan Nayak, Xiangru Jian, Kevin Qinghong Lin, Juan A Rodriguez, Montek Kalsi, Rabiul Awal, Nicolas Chapados, M Tamer Özsu, Aishwarya Agrawal, David Vazquez, et al. Ui-vision: A desktop-centric gui benchmark for visual perception and interaction. *arXiv preprint arXiv:2503.15661*, 2025.

OpenAI. GPT Image 2. https://openai.com/index/introducing-chatgpt-images-2-0/, 2026. AI image generation and editing tool.

Siyou Pei, David Kim, Alex Olwal, Yang Zhang, and Ruofei Du. UI Mobility Control in XR: Switching UI Positionings between Static, Dynamic, and Self Entities. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, pp. 1–12. ACM, May 2024. doi: 10.1145/3613904.3642220. URL http://dx.doi.org/10.1145/3613904.3642220.

Taiying Peng, Jiacheng Hua, Miao Liu, and Feng Lu. In the eye of MLLM: Benchmarking egocentric video intent understanding with gaze-guided prompting. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=YLxwj5YLb7.

Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, et al. Ui-tars: Pioneering automated gui interaction with native agents. *arXiv preprint arXiv:2501.12326*, 2025.

Qwen Team. Qwen3.6-Plus: Towards real world agents, April 2026. URL https://qwen.ai/blog?id=qwen3.6.

Christopher Rawles, Sarah Clinckemaillie, Yifan Chang, Jonathan Waltz, Gabrielle Lau, Marybeth Fair, Alice Li, William Bishop, Wei Li, Folawiyo Campbell-Ajala, et al. Androidworld: A dynamic benchmarking environment for autonomous agents. *arXiv preprint arXiv:2405.14573*, 2024.

ByteDance Seed. Ui-tars-1.5. https://seed-tars.com/1.5, 2025.

Jingyu Shi, Rahul Jain, Seunggeun Chi, Hyungjun Doh, Hyung-gun Chi, Alexander J. Quinn, and Karthik Ramani. CARING-AI: Towards Authoring Context-aware Augmented Reality INstruction through Generative Artificial Intelligence. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, pp. 1–23. ACM, April 2025. doi: 10.1145/3706598.3713348. URL http://dx.doi.org/10.1145/3706598.3713348.

Fatemeh Shiri, Xiao-Yu Guo, Mona Golestan Far, Xin Yu, Reza Haf, and Yuan-Fang Li. An empirical analysis on spatial reasoning capabilities of large multimodal models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 21440–21455, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1195. URL https://aclanthology.org/2024.emnlp-main.1195/.

Fei Tang, Zhangxuan Gu, Zhengxi Lu, Xuyang Liu, Shuheng Shen, Changhua Meng, Wen Wang, Wenqi Zhang, Yongliang Shen, Weiming Lu, Jun Xiao, and Yueting Zhuang. Gui-g^2: Gaussian reward modeling for gui grounding, 2025a. URL https://arxiv.org/abs/2507.15846.

Liang Tang, Shuxian Li, Yuhao Cheng, Yukang Huo, Zhepeng Wang, Yiqiang Yan, Kaer Huang, Yanzhe Jing, and Tiaonan Duan. Sea: Self-evolution agent with step-wise reward for computer use. *arXiv preprint arXiv:2508.04037*, 2025b.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, Congcong Wang, Dehao Zhang, Dikang Du, Dongliang Wang, Enming Yuan, Enzhe Lu, Fang Li, Flood Sung, Guangda Wei, Guokun Lai, Han Zhu, Hao Ding, Hao Hu, Hao Yang, Hao Zhang, Haoning Wu, Haotian Yao, Haoyu Lu, Heng Wang, Hongcheng Gao, Huabin Zheng, Jiaming Li, Jianlin Su, Jianzhou Wang, Jiaqi Deng, Jiezhong Qiu, Jin Xie, Jinhong Wang, Jingyuan Liu, Junjie Yan, Kun Ouyang, Liang Chen, Lin Sui, Longhui Yu, Mengfan Dong, Mengnan Dong, Nuo Xu, Pengyu Cheng, Qizheng Gu, Runjie Zhou, Shaowei Liu, Sihan Cao, Tao Yu, Tianhui Song, Tongtong Bai, Wei Song, Weiran He, Weixiao Huang, Weixin Xu, Xiaokun Yuan, Xingcheng Yao, Xingzhe Wu, Xinxing Zu, Xinyu Zhou, Xinyuan Wang, Y. Charles, Yan Zhong, Yang Li, Yangyang Hu, Yanru Chen, Yejie Wang, Yibo Liu, Yibo Miao, Yidao Qin, Yimin Chen, Yiping Bao, Yiqin Wang, Yongsheng Kang, Yuanxin Liu, Yulun Du, Yuxin Wu, Yuzhi Wang, Yuzi Yan, Zaida Zhou, Zhaowei Li, Zhejun Jiang, Zheng Zhang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Zijia Zhao, and Ziwei Chen. Kimi-VL technical report, 2025. URL https://arxiv.org/abs/2504.07491.

Venus Team, Changlong Gao, Zhangxuan Gu, Yulin Liu, Xinyu Qiu, Shuheng Shen, Yue Wen, Tianyu Xia, Zhenyu Xu, Zhengwen Zeng, Beitong Zhou, Xingran Zhou, Weizhi Chen, Sunhao Dai, Jingya Dou, Yichen Gong, Yuan Guo, Zhenlin Guo, Feng Li, Qian Li, Jinzhen Lin, Yuqi Zhou, Linchao Zhu, Liang Chen, Zhenyu Guo, Changhua Meng, and Weiqiang Wang. Ui-venus-1.5 technical report, 2026. URL https://arxiv.org/abs/2602.09082.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. Label Studio: Data labeling software, 2020-2025. URL https://github.com/HumanSignal/label-studio. Open source software available from https://github.com/HumanSignal/label-studio.

Tram Thi Minh Tran, Shane Brown, Oliver Weidlich, Soojeong Yoo, and Callum Parker. Wearable AR in Everyday Contexts: Insights from a Digital Ethnography of YouTube Videos. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, pp. 1–18. ACM, April 2025. doi: 10.1145/3706598.3713572. URL http://dx.doi.org/10.1145/3706598.3713572.

Ching-Yi Tsai, Nicole Tacconi, Andrew D. Wilson, and Parastoo Abtahi. Uncertain pointer: Situated feedforward visualizations for ambiguity-aware ar target selection. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, pp. 1–31. ACM, April 2026. doi: 10.1145/3772318.3790329. URL http://dx.doi.org/10.1145/3772318.3790329.

Haoming Wang, Haoyang Zou, Huatong Song, Jiazhan Feng, Junjie Fang, Junting Lu, Longxiang Liu, Qinyu Luo, Shihao Liang, Shijue Huang, et al. Ui-tars-2 technical report: Advancing gui agent with multi-turn reinforcement learning. *arXiv preprint arXiv:2509.02544*, 2025a.

Junyang Wang, Haiyang Xu, Jiabo Ye, Ming Yan, Weizhou Shen, Ji Zhang, Fei Huang, and Jitao Sang. Mobile-agent: Autonomous multi-modal mobile device agent with visual perception. *arXiv preprint arXiv:2401.16158*, 2024.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025b. URL https://arxiv.org/abs/2508.18265.

Xinyuan Wang, Bowen Wang, Dunjie Lu, Junlin Yang, Tianbao Xie, Junli Wang, Jiaqi Deng, Xiaole Guo, Yiheng Xu, Chen Henry Wu, Zhennan Shen, Zhuokai Li, Ryan Li, Xiaochuan Li, Junda Chen, Boyuan Zheng, Peihang Li, Fangyu Lei, Ruisheng Cao, Yeqiao Fu, Dongchan Shin, Martin Shin, Jiarui Hu, Yuyan Wang, Jixuan Chen, Yuxiao Ye, Danyang Zhang, Dikang Du, Hao Hu, Huarong Chen, Zaida Zhou, Haotian Yao, Ziwei Chen, Qizheng Gu, Yipu Wang, Heng Wang, Diyi Yang, Victor Zhong, Flood Sung, Y. Charles, Zhilin Yang, and Tao Yu. Opencua: Open foundations for computer-use agents, 2025c. URL https://arxiv.org/abs/2508.09123.

Xuehui Wang, Zhenyu Wu, JingJing Xie, Zichen Ding, Bowen Yang, Zehao Li, Zhaoyang Liu, Qingyun Li, Xuan Dong, Zhe Chen, et al. Mmbench-gui: Hierarchical multi-platform evaluation framework for gui agents. *arXiv preprint arXiv:2507.19478*, 2025d.

Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*, 2025e.

Zhenhailong Wang, Haiyang Xu, Junyang Wang, Xi Zhang, Ming Yan, Ji Zhang, Fei Huang, and Heng Ji. Mobile-agent-e: Self-evolving mobile assistant for complex tasks. *arXiv preprint arXiv:2501.11733*, 2025f.

Qianhui Wu, Kanzhi Cheng, Rui Yang, Chaoyun Zhang, Jianwei Yang, Huiqiang Jiang, Jian Mu, Baolin Peng, Bo Qiao, Reuben Tan, et al. Gui-actor: Coordinate-free visual grounding for gui agents. *arXiv preprint arXiv:2506.03143*, 2025.

Zhiyong Wu, Zhenyu Wu, Fangzhi Xu, Yian Wang, Qiushi Sun, Chengyou Jia, Kanzhi Cheng, Zichen Ding, Liheng Chen, Paul Pu Liang, et al. Os-atlas: A foundation action model for generalist gui agents. *arXiv preprint arXiv:2410.23218*, 2024.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh J Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37:52040–52094, 2024.

Tianbao Xie, Jiaqi Deng, Xiaochuan Li, Junlin Yang, Haoyuan Wu, Jixuan Chen, Wenjing Hu, Xinyuan Wang, Yuhui Xu, Zekun Wang, et al. Scaling computer-use grounding via user interface decomposition and synthesis. *arXiv preprint arXiv:2505.13227*, 2025.

Yiheng Xu, Zekun Wang, Junli Wang, Dunjie Lu, Tianbao Xie, Amrita Saha, Doyen Sahoo, Tao Yu, and Caiming Xiong. Aguvis: Unified pure vision agents for autonomous gui interaction. *arXiv preprint arXiv:2412.04454*, 2024.

Yan Yang, Dongxu Li, Yutong Dai, Yuhao Yang, Ziyang Luo, Zirui Zhao, Zhiyuan Hu, Junzhe Huang, Amrita Saha, Zeyuan Chen, Ran Xu, Liyuan Pan, Silvio Savarese, Caiming Xiong, and Junnan Li. Gta1: Gui test-time scaling agent, 2025. URL https://arxiv.org/abs/2507.05791.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, Bokai Xu, Junbo Cui, Yingjing Xu, Liqing Ruan, Luoyuan Zhang, Hanyu Liu, Jingkun Tang, Hongyuan Liu, Qining Guo, Wenhao Hu, Bingxiang He, Jie Zhou, Jie Cai, Ji Qi, Zonghao Guo, Chi Chen, Guoyang Zeng, Yuxuan Li, Ganqu Cui, Ning Ding, Xu Han, Yuan Yao, Zhiyuan Liu, and Maosong Sun. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe, 2025. URL https://arxiv.org/abs/2509.18154.

Borui Zhang, Bo Zhang, Bo Wang, Wenzhao Zheng, Yuhao Cheng, Liang Tang, Yiqiang Yan, Jie Zhou, and Jiwen Lu. Bami: Training-free bias mitigation in gui grounding, 2026. URL https://arxiv.org/abs/2605.06664.

Yunzhu Zhang, Zeyu Pan, Zhengwen Zeng, Shuheng Shen, Changhua Meng, and Linchao Zhu. Mvp: Multiple view prediction improves gui grounding. *arXiv preprint arXiv:2512.08529*, 2025.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

Yawei Zhao, Jiaxin Liang, Hao Li, and Pan Hui. Where Digital Meets Place: Deriving Strategies for Curating Mixed Reality Exhibitions in Public Spaces. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, pp. 1–23. ACM, April 2026. doi: 10.1145/3772318.3790893. URL http://dx.doi.org/10.1145/3772318.3790893.

Beitong Zhou, Zhexiao Huang, Yuan Guo, Zhangxuan Gu, Tianyu Xia, Zichen Luo, Fei Tang, Dehan Kong, Yanyi Shang, Suling Ou, et al. Venusbench-gd: A comprehensive multi-platform gui benchmark for diverse grounding tasks. *arXiv preprint arXiv:2512.16501*, 2025a.

Hanzhang Zhou, Xu Zhang, Panrong Tong, Jianan Zhang, Liangyu Chen, Quyu Kong, Chenglin Cai, Chen Liu, Yue Wang, Jingren Zhou, et al. Mai-ui technical report: Real-world centric foundation gui agents. *arXiv preprint arXiv:2512.22047*, 2025b.

Qiushi Zhou, Jean Paul Vera Soto, Zhongyi Bai, Mark Parent, Kashyap Todi, Tanya R. Jonker, and Eduardo Velloso. A Probabilistic Approach to Understanding User Preferences for Adaptive Placement of AR Interfaces in Different Physical Environments. *IEEE Transactions on Visualization and Computer Graphics*, 32(7):5483–5500, July 2026. ISSN 2160-9306. doi: 10.1109/tvcg.2026.3667949. URL http://dx.doi.org/10.1109/TVCG.2026.3667949.

Sheng Zhou, Junbin Xiao, Qingyun Li, Yicong Li, Xun Yang, Dan Guo, Meng Wang, Tat-Seng Chua, and Angela Yao. Egotextvqa: Towards egocentric scene-text aware video question answering. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3363–3373, 2025c. URL https://api.semanticscholar.org/CorpusID:276258564.

Qian Zhu, Zhuo Wang, Wei Zeng, Wai Tong, Weiyue Lin, and Xiaojuan Ma. Make Interaction Situated: Designing User Acceptable Interaction for Situated Visualization in Public Environments. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, pp. 1–21. ACM, May 2024. doi: 10.1145/3613904.3642049. URL http://dx.doi.org/10.1145/3613904.3642049.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pp. 2165–2183. PMLR, 2023.

Stefanie Zollmann, Tobias Langlotz, Raphael Grasset, Wei Hong Lo, Shohei Mori, and Holger Regenbrecht. Visualization Techniques in Augmented Reality: A Taxonomy, Methods and Patterns. *IEEE Transactions on Visualization and Computer Graphics*, 27(9):3808–3825, September 2021. ISSN 2160-9306. doi: 10.1109/tvcg.2020.2986247. URL http://dx.doi.org/10.1109/TVCG.2020.2986247.

This is the supplementary material for the paper "EgoXR-GUI: Benchmarking GUI Grounding in Physical–Digital Extended Reality". We organize the content as follows:

A – Extended Dataset Construction Details

B – Additional Results and Analysis

C – Downstream Application: Prototyping an XR GUI Agent

D – Additional Visualization

E – Experimental and Implementation Details

F – Limitations and Future Works

G – Broader Impact

A Extended Dataset Construction Details

A.1 Data Collection & Privacy Protocol

To ensure the ethical integrity of the egocentric data collection process, we implemented rigorous protocols to protect the privacy of both data collectors and third parties. Our GUI grounding dataset comprises screen recordings captured from extended reality (XR) devices in various private and public environments. We obtained informed consent from all data collectors prior to the study, ensuring they are fully aware of the recording scope, and we strictly prohibit the storage or transmission of any sensitive personal identifiers. Regarding the privacy of bystanders and non-participants, we apply automated blurring algorithms to mask all identifiable features, such as faces encountered in the background of the XR interface. These measures ensure that the resulting dataset remains anonymized and compliant with standard privacy regulations.

A.2 Annotation Guidelines

To construct the benchmark for evaluating the GUI grounding capabilities of various models, we perform bounding box annotation on the collected screenshots using the Label Studio platform. Our annotation criterion requires that each bounding box encompasses the entire region of the correct UI element to a reasonable extent. Instructions explicitly define the grounding scope by specifying the target window. The standard templates for these instructions typically follow a structured format, such as opening a specific page of an application within a designated window or performing a particular operation inside a given window.

Annotation quality audit. The two annotators, denoted A1 and A2, are members of the author team. Participants recorded their task intentions during data collection; the annotators used these intentions and the screenshot context to formulate standardized instructions and label the target boxes. Each annotator then reviewed 100 annotations produced by the other. One original participant additionally reviewed 100 samples originating from that participant's data to check that the standardized instructions preserved the intended goals.

We checked five criteria: whether the task fits the scene and activity (scene fit), whether it satisfies its grounding-task definition (task-type fit), whether the instruction follows the annotation template (format), whether the box covers the intended UI element (box), and whether the annotation preserves the originating task goal (user intent). Table 5 reports the criterion-level pass rates for each subset. The subsets are reported separately and may overlap.

A.3 Comprehensive Dataset Statistics

In addition to the primary evaluation, we provide more granular statistics to characterize the proposed benchmark, including **Instruction Length**, **Element Size**, and the **Number of UI Windows** per sample. We also report the corresponding grounding accuracy of state-of-the-art (SOTA) models across these dimensions. As illustrated in Fig. 5 and Fig 6, these results offer a comprehensive view

Table 5: Quality audit of real benchmark annotations. Each row contains 100 audited annotations; entries are criterion-level pass rates (%).

Audited subset	Scene	Task type	Format	Box	Intent
A1 annotations, checked by A2	100	100	99	100	100
A2 annotations, checked by A1	100	99	100	100	100
Participant's samples, checked by that participant	100	100	100	100	100

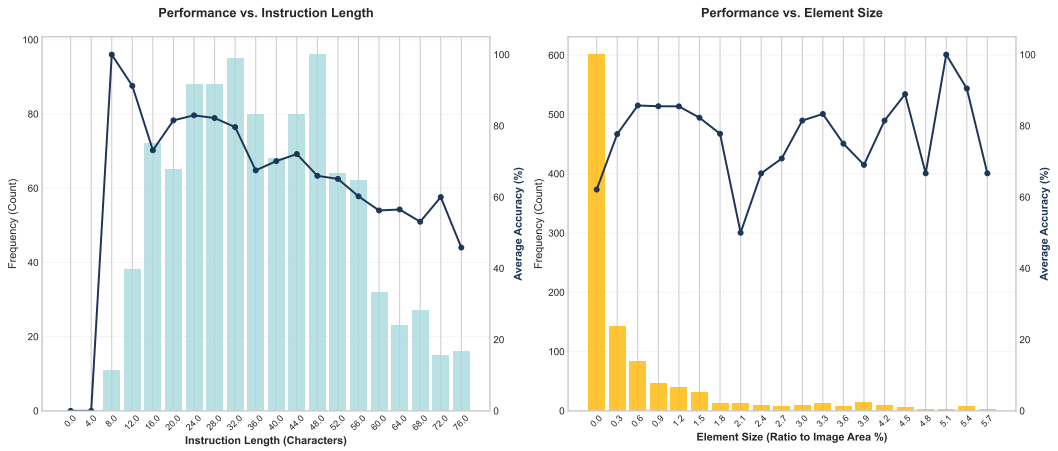


Figure 5: **Quantitative analysis of grounding performance across instruction length and element size.** The bar charts represent the frequency distribution (count) of samples, while the line graphs illustrate the average accuracy of the baseline model. (Left) **Performance vs. instruction length**: The impact of linguistic complexity (character count) on grounding success. (Right) **Performance vs. element size**: The relationship between UI element scale (ratio to image area %) and localization accuracy, highlighting that accurate localization of extremely small elements remains a significant challenge in complex XR backgrounds.

of model robustness and highlight specific challenges, such as accurately localizing extremely small elements within complex XR environments.

A.4 XR Design Context and Interface Coverage

The benchmark organizes XR GUI grounding along three complementary dimensions: the visual and contextual challenges encountered by the model, the activities in which an interface is used, and the interface elements selected to accomplish a task. Table 6 relates these dimensions to prior XR research. We synthesize the four challenge dimensions for GUI grounding, while the three task categories distinguish the evidence required to identify the target: explicit UI attributes, spatial relations, or physical-scene semantics and task intent.

Our collection hierarchy comprises six usage domains, ten places, and more than twenty activities. For each activity, the participant's goal determines the relevant interface and target element. Table 7 groups the 47 collected applications and interfaces into twelve functional categories. These counts describe interface coverage rather than the frequency of task samples. They include native applications, web or mini-program interfaces, and device-control panels. Games are outside the current collection scope.

Task-name compatibility. Explicit Grounding is the task previously named Direct Grounding. Existing data and evaluation code may retain the legacy `direct` tag; both names refer to the same samples, target annotations, and evaluation protocol.

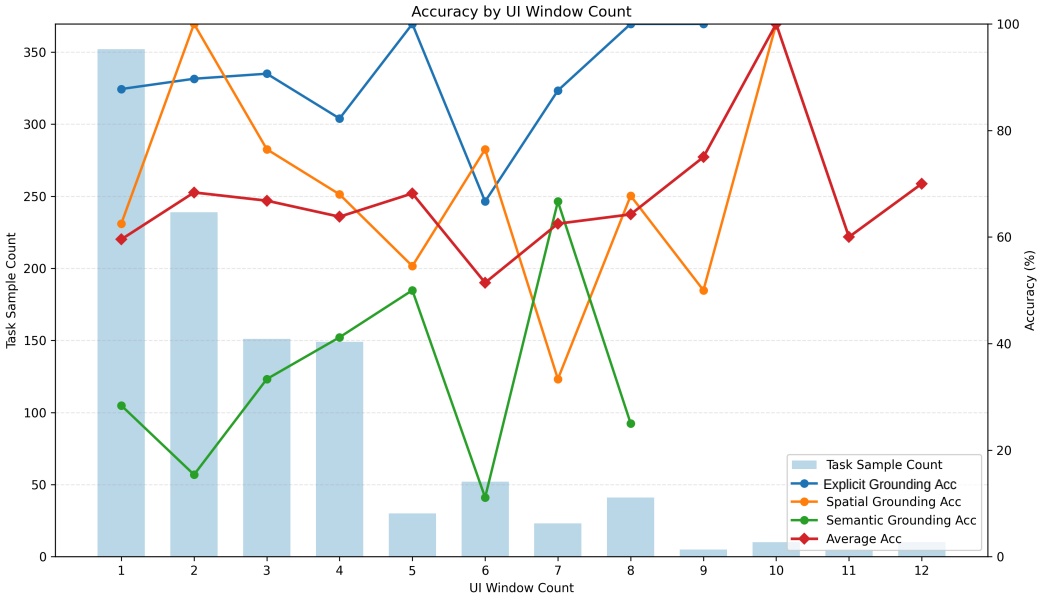


Figure 6: **Analysis of accuracy across different UI window counts.** The bar chart represents the task sample count (left axis), while the line plots illustrate the accuracy for explicit, spatial, and semantic grounding tasks alongside the average accuracy (right axis). Performance fluctuations are observed as the complexity of window counts increases.

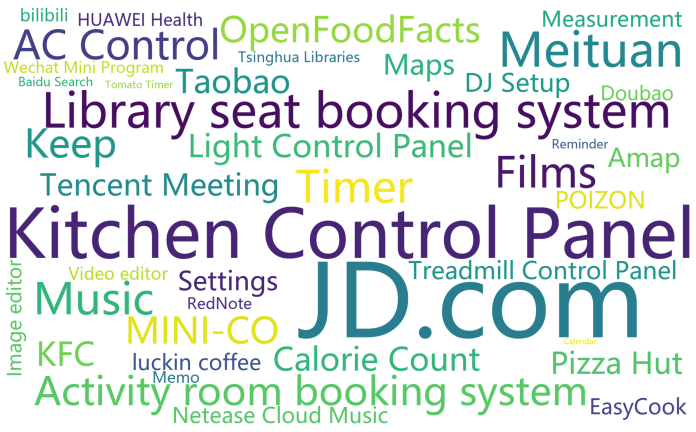


Figure 7: **Visualization of task scenarios and application domains.** The word cloud illustrates the diversity of our dataset, covering various GUI applications, ranging from popular mobile apps to smart home control panels.

Table 6: Connections between prior XR research and the design dimensions of EgoXR-GUI.

Design dimension	Prior XR evidence	Connection to EgoXR-GUI
Visibility and rendering	Real–virtual depth ordering, occlusion handling, and AR visualization techniques (Macedo & Apolinario, 2023; Zollmann et al., 2021).	Perceptual perturbation and XR-rendering complexity in the captured observation.
Virtual–physical distinction	Separate descriptions of physical scenes and projected content in projection-based AR (Cao et al., 2026).	Distinguishing interface content from physical scene entities in Semantic Grounding.
Spatial reference	Interface anchoring and contextual placement (Pei et al., 2024; Li et al., 2024b); layout preferences emphasizing structure and consistency across physical environments (Zhou et al., 2026).	Spatial Grounding relative to physical entities and viewpoint-dependent interface layouts.
Situated intent	Contextual assistance and spoken references (Li et al., 2025a; Kim et al., 2026b); spatial-memory-based intent resolution (Kim et al., 2026a) and context-aware assistance delivery (Lee et al., 2025).	Semantic Grounding from scene context and user intent within a single observation.
Everyday activities	Everyday media, productivity, and wellness uses; personalized interfaces and glanceable information access (Tran et al., 2025; Cho et al., 2024; Lu & Bowman, 2021).	Scenario–place–activity–task organization for daily XR use.
Contextual authoring	Authoring AR instructions by combining task content with spatial and temporal context (Shi et al., 2025).	Creation activities and the relationship between user goals and physical context.
Interface composition	Widget functions, UI element types, and placement in personalized XR layouts (Cho et al., 2024).	Activity-driven selection of applications, interfaces, and actionable target elements.
Target disambiguation and selection	Target density, visibility, and ambiguity-aware representations (Tsai et al., 2026); a taxonomy of selection tasks in consumer VR applications (Maslych et al., 2024).	Perceptual disambiguation and organization of actionable target elements by task.
Public-place context	Situated interaction acceptability (Zhu et al., 2024) and place-specific content organization in MR exhibitions (Zhao et al., 2026).	Background on relationships between activities, interfaces, and place.

Table 7: Functional coverage of the 47 applications and interfaces in EgoXR-GUI.

Functional category	Count	Included interfaces
Food and dining	8	McDonald's, Luckin Coffee, KFC Ordering, Pizza Hut Ordering, Meituan, EasyCook, OpenFoodFacts, Taobao Instant Commerce
Productivity and office	6	Calendar, Reminders, Notes, Pomodoro Timer, Timer, Tencent Meeting
Learning and information access	6	Tsinghua Library, Library Seat Reservation, Activity Room Reservation, Baidu Search, Doubao, AI
Media and entertainment	5	Bilibili, NetEase Cloud Music, Music, Video Player, DJ Console
System and utilities	5	Time, Weather, Settings, Measure, Stocks
Sports and health	4	Keep, Huawei Health, Calorie Logging, Treadmill Control
Smart home and device control	3	Air Conditioner Control, Lighting Control Panel, Kitchen Control
Shopping	3	JD.com, Dewu, Taobao
Travel and navigation	2	Amap, Maps
Creation and editing	2	Video Editing, Image Editing
Entry points and platforms	2	Xiaoka, WeChat Mini-Program Selector
Social and content	1	Xiaohongshu
Total	47	

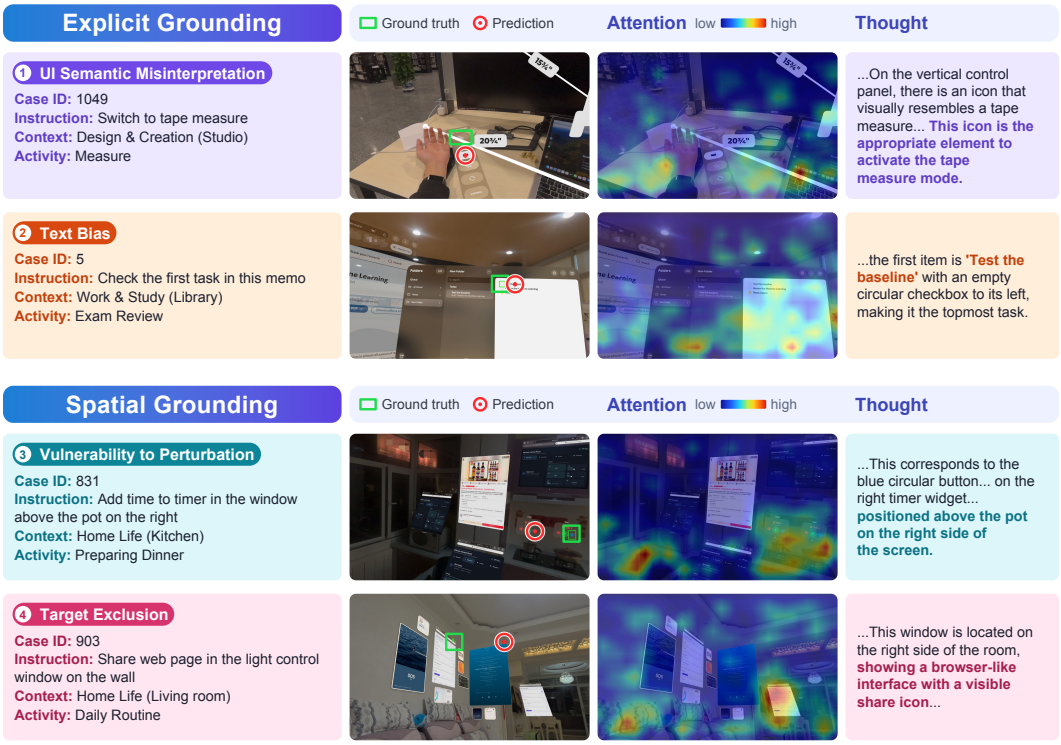


Figure 8: **Error case analysis of explicit and spatial grounding.** This figure illustrates representative failure modes using MAI-UI 8B. In **Explicit Grounding**, the model exhibits *UI Semantic Misinterpretation* (Case 1049) due to recognition bias in complex interfaces, and *Text Bias* (Case 5) resulting from an inherent inductive preference for text elements. In **Spatial Grounding**, the heatmaps reveal *Vulnerability to Perturbation* (Case 831), where distractors interfere with grounding results, and *Target Exclusion* (Case 903), where the model fails to resolve spatial relationships within multi-window scenarios.

B Additional Results and Analysis

B.1 Error Analysis on Explicit Grounding

Model errors in Explicit Grounding primarily stem from two factors.

• *UI Semantic Misinterpretation*: Recognition bias and comprehension ambiguity in complex interfaces, as shown in the first case of Fig. 8 (Case 1049). This manifests as confusion between similar UI elements, incorrect functional interpretation, or complete failure to detect targets.

• *Text Bias*: An inductive bias toward text elements, where predicted coordinates land on text adjacent to the ground truth icon, as shown in the second case of Fig. 8 (Case 5). This phenomenon often originates from the inherent text-centric bias of VLMs.

Such errors typically result from knowledge gaps and reduced generalization due to insufficient training data. These issues have been validated in prior benchmarks. However, following advancements in foundational VLMs and training methodologies, recent models achieve high accuracy on these tasks.

B.2 Error Analysis on Spatial Grounding

Analysis reveals that the significantly lower accuracy of most grounding models in Spatial Grounding compared to Explicit Grounding primarily stems from their inability to resolve spatial relationships within multi-window scenarios. Based on attention heatmaps, we identify two distinct error patterns.

Table 8: Ablation study of Qwen3-VL Thinking versions. Arrows indicate changes relative to the corresponding non-Thinking model; a dash denotes no change.

Models	Explicit Grounding			Spatial Grounding			Semantic Grounding			Overall		
	text	icon	avg	text	icon	avg	text	icon	avg	text	icon	avg
Qwen3-VL-2B	80.70	55.72	71.39	27.27	36.36	32.39	15.50	11.43	14.63	61.69	47.16	56.18
+ Thinking	63.82↓	45.02↓	56.81↓	19.48↓	30.30↓	25.57↓	20.93↑	17.14↑	20.12↑	50.23↓	39.01↓	45.97↓
Qwen3-VL-8B	90.13	73.80	84.04	49.35	52.53	51.14	24.03	20.00	23.17	72.55	63.95	69.29
+ Thinking	84.65↓	67.16↓	78.13↓	36.36↓	43.43↓	40.34↓	41.86↑	20.00–	37.20↑	70.74↓	57.28↓	65.64↓
Qwen3-VL-32B	90.35	62.36	79.92	58.44	53.54	55.68	46.51	20.00	40.85	78.13	56.54	69.94
+ Thinking	86.84↓	70.85↑	80.88↑	37.66↓	58.59↑	49.43↓	55.04↑	37.14↑	51.22↑	74.96↓	64.94↑	71.16↑
Qwen3-VL-30B-A3B	86.84	71.96	81.29	46.75	43.43	44.89	20.93	17.14	20.12	69.38	60.25	65.92
+ Thinking	88.16↑	69.00↓	81.02↓	41.56↓	47.47↑	44.89–	46.51↑	28.57↑	42.68↑	74.66↑	60.25–	69.19↑

• *Vulnerability to Perturbation*: The attention region covers both the target and the misleading windows, yet the model remains susceptible to interference, leading to an incorrect grounding result. The third case of Fig. 8 (Case 831) illustrates this failure mode.

• *Target Exclusion*: The model's attention region fails to encompass the target UI window, instead concentrating on other misleading distractors, as shown in the last case of Fig. 8 (Case 903).

These observations collectively validate that current grounding models lack sufficient spatial reasoning capabilities and struggle to accurately interpret relative spatial positions described in instructions or depicted in images.

B.3 Agentic Inference Frameworks

To bridge the gap between high-level reasoning and fine-grained localization, we investigated three agentic inference frameworks. These frameworks utilize a dual-model pipeline: a high-level reasoning model (e.g., Doubao-seed-2.0 or Seed 1.6 VL) for semantic understanding and a specialized grounding model (MAI-UI 8B) for final coordinate prediction.

1. **Color BBox Mark**: This framework leverages visual prompting to assist grounding. The reasoning model identifies task-relevant physical objects and UI elements, marking them with color-coded bounding boxes directly on the raw image. These marked images, along with corresponding semantic descriptions, are then passed to the grounding model. By grounding within a visually structured context, the system synthesizes superior semantic comprehension with specialized localization precision.
2. **Agent-S3**: Based on the state-of-the-art framework for OS-World (Gonzalez-Pumariega et al., 2025), we implement a hierarchical pipeline. The high-level model performs semantic reasoning to translate ambiguous or abstract user queries into explicit grounding instructions (e.g., "Click the volume button in the settings window"). The grounding model then executes these simplified instructions to generate the final coordinates.
3. **Task Router**: Building upon the Agent-S3 framework, we introduce an adaptive routing mechanism using an additional lightweight VLM. This router evaluates each incoming instruction and visual context to classify the task type. The complex (and computationally expensive) reasoning pathway is activated exclusively for tasks identified as *Semantic Grounding*. For *Explicit* and *Spatial* grounding, the router bypasses the reasoning model and triggers direct prediction, reserving additional semantic reasoning for context-dependent queries.

As shown in Tab. 3 (Main Paper), the Task Router strategy achieves the highest average accuracy among the compared frameworks by selectively applying the semantic reasoning pathway.

B.4 Ablation Study of Qwen3-VL Thinking Version

We conducted an ablation study across various parameter scales of the Qwen3-VL series to evaluate their Thinking versions. The experimental results in Table 8 show that the Thinking versions improve

average Semantic Grounding accuracy at all four scales, reaching 20.12%, 37.20%, 51.22%, and 42.68% for 2B, 8B, 32B, and 30B-A3B, respectively.

Conversely, we observed that overall accuracy decreases for the smaller 2B and 8B models, from 56.18% to 45.97% and from 69.29% to 65.64%, respectively, while increasing for 32B and 30B-A3B. Average Spatial Grounding accuracy decreases for 2B, 8B, and 32B and remains unchanged for 30B-A3B, indicating that the semantic gains do not consistently extend to spatial localization.

B.5 Effeciency and Results of Chinese Instruction

Table 9: Experimental results on different task types. We evaluate 28 models across five categories on three grounding tasks. All numbers are reported in percentage (%).

Models	Explicit Grounding			Spatial Grounding			Semantic Grounding			Overall			Efficiency
	text	icon	avg	text	icon	avg	text	icon	avg	text	icon	avg	Avg. Time / Sample(s)
Closed-source Models													
Gemini 3	33.77	32.10	33.15	20.78	20.20	20.45	39.53	20.00	35.37	33.48	28.15	31.46	12.46
Seed1.6 VL	74.34	68.27	72.08	32.47	57.58	46.59	49.61	31.43	45.73	64.71	62.47	63.86	12.98
General Open-source VLMs													
Qwen2.5-VL-7B-Instruct	73.46	48.71	64.24	29.87	34.34	32.39	24.81	5.71	20.73	58.82	41.48	52.25	5.13
Qwen3-VL-2B	80.70	52.40	70.15	23.38	34.34	29.55	13.95	11.43	13.41	60.94	44.44	54.68	4.23
Qwen3-VL-8B	89.25	71.96	82.81	53.25	49.49	51.14	27.91	14.29	25.00	73.15	61.48	68.73	6.04
Qwen3-VL-32B	88.82	61.25	78.54	58.44	53.54	55.68	48.84	17.14	42.07	77.53	55.56	69.19	11.95
Qwen3-VL-30B-A3B	88.82	71.59	82.39	53.25	55.56	54.55	27.13	17.14	25.00	72.70	62.96	69.01	12.31
InternVL3.5-8B	58.99	38.75	51.44	22.08	31.31	27.27	16.28	2.86	13.41	46.30	33.83	41.57	1.48
InternVL3.5-30B-A3B	64.25	43.91	56.67	22.08	27.27	25.00	22.48	8.57	19.51	51.13	36.79	45.69	4.88
MiniCPM-V4.5-8B	49.78	25.09	40.58	6.49	7.07	6.82	15.50	2.86	12.80	38.01	18.77	30.71	1.03
GUI-specific Models (SFT)													
OS-Atlas-7B	56.36	36.53	48.97	16.88	19.19	18.18	11.63	2.86	9.76	42.99	29.38	37.83	1.19
OpenCUA-7B	44.30	47.60	45.53	14.29	40.40	28.98	5.43	8.57	6.10	33.18	42.47	36.70	4.83
OpenCUA-32B	69.30	60.52	66.02	35.06	39.39	37.50	14.73	8.57	13.41	54.75	50.86	53.28	11.41
OpenCUA-72B	8.55	11.07	9.49	15.58	27.27	22.16	0.00	0.00	0.00	7.69	14.07	10.11	33.33
GUI-specific Models (RL)													
Holo1.5-72B	88.16	73.06	82.53	54.55	67.68	61.93	41.09	11.43	34.76	75.11	66.42	71.82	4.11
Holo2-4B	79.39	57.56	71.25	37.66	40.40	39.20	16.28	8.57	14.63	62.29	49.14	57.30	2.34
Holo2-8B	83.33	60.15	74.69	32.47	37.37	35.23	18.60	5.71	15.85	64.86	49.88	59.18	3.43
Holo2-30B-A3B	90.35	73.43	84.04	57.14	66.67	62.50	21.71	20.00	21.34	73.15	67.16	70.88	8.80
UI-Venus 1.5-30B-A3B	91.67	80.07	87.35	64.94	73.74	69.89	31.78	11.43	27.44	76.92	72.59	75.28	7.95
MAI-UI-8B	91.01	82.66	87.90	64.94	59.60	61.93	40.31	34.29	39.02	78.13	72.84	76.12	3.55

We provide the effeciency and results of Chinese instruction in Tab. 9

B.6 Challenge Factors in the Error Cases

Table 10 connects the error cases in Figs. 4 and 8 to the four challenge dimensions introduced in Section 1. Several factors can occur in the same example: target exclusion combines window rendering, spatial relations, and semantic disambiguation, whereas physical mislocalization involves identifying the relevant scene entity and relating it to the instruction. The table provides a qualitative reading of these examples; check marks denote contributing factors rather than separate error counts.

B.7 Cross-device Semantic Grounding Study

We recaptured Semantic Grounding cases on Pico 4 Ultra to examine whether failures observed on Apple Vision Pro also occur on another headset. The study uses 18 matched pairs selected from UI-Venus-1.5-8B failures on the Apple data, with corresponding instructions and intended targets on Pico. The matching records contain 18 distinct Pico images and 13 distinct Apple images across four capture locations: supermarket (11 pairs), restaurant (3), meeting room (3), and library (1).

We compute cosine similarity between L2-normalized image embeddings from `openai/clip-vit-base-patch32`. Table 11 compares the matched pairs with controls from the same or different capture locations. For each control, we first average a source image's

Table 10: Challenge factors illustrated by the existing error cases. P: perceptual perturbation; R: XR-rendering complexity; Sp: spatial uncertainty; Se: semantic ambiguity.

Task and observed failure mode	P	R	Sp	Se
Explicit: UI semantic misinterpretation (Case 1049)	✓	✓		✓
Explicit: text bias (Case 5)		✓		✓
Spatial: vulnerability to perturbation (Case 831)	✓	✓	✓	
Spatial: target exclusion (Case 903)		✓	✓	✓
Semantic: incorrect UI binding (Case 166)		✓		✓
Semantic: semantic mapping failure (Case 207)	✓			✓
Semantic: physical mislocalization (Case 281)	✓		✓	✓
Semantic: complete physical distraction (Case 242)	✓			✓

Table 11: CLIP image similarity in the cross-device study. "Within-location" and "cross-location" controls use other captures from the same or a different collection location, respectively. The matched row summarizes individual pairs; control rows summarize per-source means.

Comparison	N	Mean $\pm$ SD	Median	Range
Pico–Apple, task-matched	18	0.8024 ± 0.0466	0.8020	0.7262–0.8843
Pico–Pico, within-location	17	0.8174 ± 0.0409	0.8169	0.7642–0.8948
Pico–Apple, within-location	18	0.7581 ± 0.0455	0.7589	0.6994–0.8473
Pico–Pico, cross-location	18	0.7035 ± 0.0452	0.7183	0.6086–0.7689
Pico–Apple, cross-location	18	0.6767 ± 0.0396	0.6785	0.5959–0.7275
Apple–Apple, within-location	11	0.7557 ± 0.0411	0.7587	0.7044–0.8252
Apple–Apple, cross-location	13	0.6534 ± 0.0314	0.6678	0.6028–0.6959

similarities to eligible images in the comparison pool, and then summarize those per-source averages. Pico controls use the 18 distinct matched Pico images; Pico–Apple controls use a comparison pool of 28 Apple images grouped by location. Apple–Apple controls use the 13 distinct matched Apple images as both sources and comparison pools. Self-pairs and the matched Apple target are excluded from the corresponding within-location controls. Thus, N counts valid source-image summaries for controls, and SD is the sample standard deviation of those summaries.

Task-matched captures achieve a mean similarity of 0.8024, compared with 0.7581 for other cross-device captures within the same location and 0.6767 across locations. UI-Venus-1.5-8B remains incorrect on all 18 corresponding Pico tasks, showing that the selected Semantic Grounding failures recur under another headset's rendering. This is a diagnostic study of selected failures; CLIP characterizes global image similarity, and the selection protocol does not estimate overall cross-device accuracy.

B.8 Grounding with Gaze Coordinates

We evaluate whether gaze points collected with the screenshots help MAI-UI-8B identify the intended UI target. We add the gaze location to each prompt as a short first-person description, using either raw pixel coordinates or coordinates normalized to $[0, 1000]$. All other task definitions and accuracy criteria follow the baseline evaluation.

Gaze prompt. We prepend the same text template to the original English instruction in both settings:

```
I look at [{x}, {y}] in this image.  {instruction}
```

Let (g_x, g_y) denote the gaze point in original screenshot pixels and (W, H) the image dimensions stored in the annotation. The raw setting uses $(x, y) = (g_x, g_y)$; the normalized setting uses $x = \mathrm{round}(1000 g_x / W)$ and $y = \mathrm{round}(1000 g_y / H)$, with Python's integer rounding. Coordinates are formed before the model's image resizing. Both settings use the same system message. The full system and user-message templates are provided in Appendix E.1.

Table 12: MAI-UI-8B with gaze coordinates supplied as text. Each cell reports average/Text/Icon grounding accuracy (%). Average accuracy uses the same sample-weighted metric as Table 2.

Input	Explicit	Spatial	Semantic	Overall
Baseline	88.17/92.11/81.55	64.77/63.64/65.66	30.49/31.78/25.71	75.47/77.07/72.84
Raw gaze	78.13/81.80/71.96	59.09/54.55/62.63	29.88/33.33/17.14	67.60/69.23/64.94
Normalized gaze	82.26/87.28/73.80	64.20/61.04/66.67	31.71/32.56/28.57	71.54/73.60/68.15

As shown in Table 12, raw and normalized gaze coordinates reduce overall accuracy from 75.47% to 67.60% and 71.54%, respectively. Normalized gaze increases Semantic Grounding accuracy from 30.49% to 31.71%. Gaze reflects visual attention, which can differ from the intended action target: a user watching a video may request an action in a food-ordering interface. The mixed results motivate incorporating gaze through task-aware routing or multimodal fusion beyond the text-coordinate encoding tested here. Visible hand–object relations already provide context for instructions such as "the object I am holding"; synchronized gaze and hand trajectories would further extend this static-frame setting.

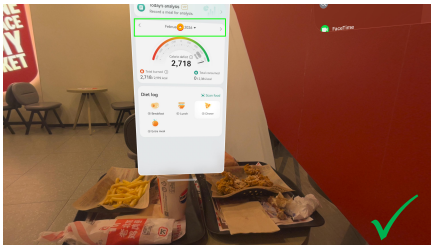
C Downstream Application: Prototyping an XR GUI Agent

While EgoXR-GUI primarily serves as an evaluation benchmark for visual grounding, the ultimate goal of accurately localizing UI elements in extended reality is to empower autonomous Embodied AI. To validate the practical significance of our benchmark, we prototyped an autonomous XR GUI Agent as a downstream application. Specifically, we developed an end-to-end agentic system based on the Agent-S3 framework, utilizing Doubao-seed-2.0 as the high-level planner and evaluator, and MAI-UI 8B—a top-performing model in our benchmark—as the grounding engine. The system was successfully deployed and evaluated on the Pico 4 Ultra platform. In the **Supplementary Video (begin from 14:27)**, we present two distinct demonstrations: a single-step semantic grounding task and a complex, multi-step explicit grounding task performed by the autonomous agent.

D Additional Visualization

D.1 More Visualization Results

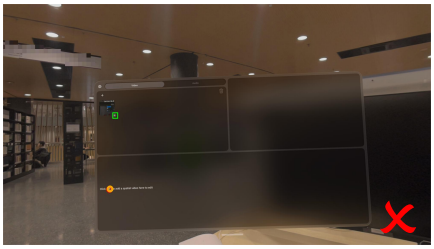
We provide additional visualization results of different tasks in our dataset in Fig 9,10,11.



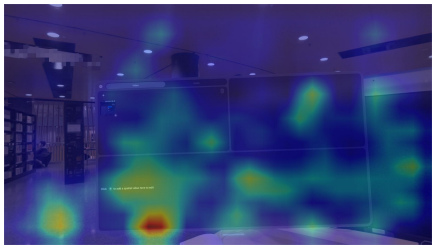
Instruction: Change record date on the APP.



Response: The instruction asks to change the record date, which corresponds to the date selector at the top center of the app interface. This element displays "February 5, 2026" with a dropdown arrow, indicating it can be opened to select a different date.



Instruction: Add video to editing space.



Response: The instruction "add video to editing space" corresponds to the small circular plus icon next to the text "Click + to add a spatial video here to edit" at the bottom left of the interface. This icon is designed to initiate adding a video, matching the instruction's intent.

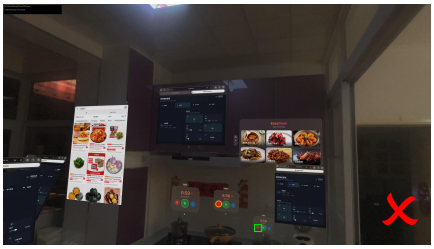
Figure 9: Qualitative results on explicit grounding tasks, with the correct sample (top) showcasing a "Sports and Health" scenario, and the incorrect sample (bottom) showcasing a "Design and Creation" scenario. In the left panel, the green box denotes the ground truth target, while the red circle and orange circle represent the predicted click points of MAI-UI 8B and UI-Venus 1.5 7B, respectively. The right panel displays the cross-modal attention heatmap and reasoning output of MAI-UI 8B model. The following two figures follow the same layout structure.



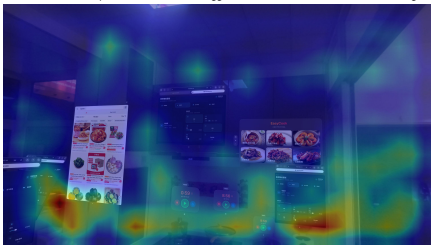
Instruction: Turn off the light power in the window next to the curtains.



Response: The instruction asks to turn off the light power in the window next to the curtains. This corresponds to the Light Control Panel interface displayed near the curtains on the left side of the room. Within this panel, the "Power" toggle is the control to switch off the light.

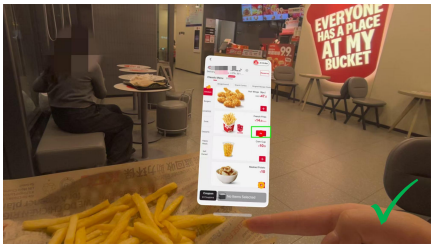


Instruction: Stop timer in the window above the pot on the right.

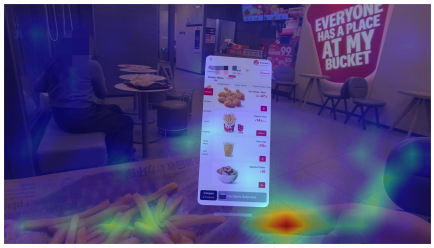


Response: The instruction asks to stop the timer in the window above the pot on the right. This corresponds to the red button on the rightmost timer interface, which is located above the pot and has a square icon indicating pause or stop.

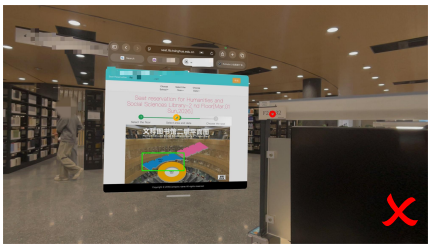
Figure 10: Qualitative results on spatial grounding tasks, with both the correct sample (top) and the incorrect sample (bottom) showcasing the "Home Life" scenario. Layout and color indicators follow the same convention as Fig. 9.



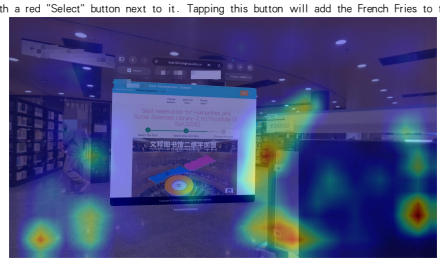
Instruction: Order the item I'm pointing to on the APP.



Response: The instruction "order the item I'm pointing to on the APP" refers to the finger pointing at the French Fries item on the menu. The French Fries item is located in the middle of the menu, with a red "Select" button next to it. Tapping this button will add the French Fries to the order.



Instruction: Select the area corresponding to the number next to the window on the seat reservation interface.



Response: The instruction asks to select the area with the number next to the window on the seat reservation interface. On the right side of the interface, there is a white sign with the text "F2A032" mounted on a beam, which is next to a window. This is the element to interact with.

Figure 11: Qualitative results on semantic grounding tasks, with the correct sample (top) showcasing a "Consumption and Entertainment" scenario, and the incorrect sample (bottom) showcasing a "Work and Study" scenario. Layout details are consistent with Fig. 9.

E Experimental and Implementation Details

E.1 Prompt Configurations

We detail the specific prompt templates used for evaluating various models.

Prompt of CogAgent

```
Task: instruction annotation History steps: WIN(Answer in
Action-Operation format.)
```

Prompt of Holo 1.5

```
Localize an element on the GUI image according to the provided
target and output a click position.
    • Only output the click position, do not output any other text.
    • The click position should be in the format 'Click (x, y)'
      with x: num pixels from the left edge and y: num pixels
      from the top edge.
Your target is: instruction annotation
```

Prompt of Holo 2

```
Localize an element on the GUI image according to the provided
target and output a click position.
    • You must output a valid JSON following the format: schema
Your target is: instruction annotation
```

Prompt of InternVL

```
Please provide the bounding box coordinate of the UI element this
user instruction describes: <ref>instruction annotation</ref>.
Answer in the format of [[x1, y1, x2, y2]]
```

Prompt of MiniCPMV

```
What is the bounding box of the UI element corresponding to the
user instruction "instruction annotation"? Output in the format of
x1 x2 y1 y2.
```

Prompt of OpenCUA

```
System Prompt: You are a GUI agent. You are given a task and
a screenshot of the screen. You need to perform a series of
pyautogui actions to complete the task.
User Prompt:
{instruction annotation}
```

Prompt of OS-Atlas-7B

```
(ground only positive) In this UI screenshot, what is the
position of the element corresponding to the command "instruction
annotation" (with bbox)?
(ground allow negative)
Output the bounding box in the image corresponding to the
instruction "{instruction annotation}". If the target does not
exist, respond with "Target does not exist".
```

Prompt of Seed-1.5-VL

(positive) You are a GUI grounding agent. ## Output Format
Action: click(point='<point>x y</point>') ## Rules

- Use pixel coordinates in the original image (width, height).
- Output only one Action line.

User Instruction instruction annotation
(negative)
You are a GUI grounding agent. ## Output Format
Action: click(point='<point>x y</point>') # If the target does not exist, output:
Action: finished(content='negative') ## Rules

- Use pixel coordinates in the original image (width, height).
- Output only one Action line.

User Instruction
{instruction annotation}

Prompt of TianXi-7B

System Prompt: You are a helpful assistant.
User Prompt:
The image is a screenshot of a computer or mobile phone interface, with a resolution of {x}x{y}. Please provide the coordinates of the object to be operated according to the command, which is as follows: {instruction annotation}.
(prompt repeat)
Repeat the task again for you:
Please provide the coordinates of the object to be operated according to the command, which is as follows: {instruction annotation}.
You must output in the following format:
<|box_start|>(x1,y1),(x2,y2)<|box_end|>

Prompt of UIVenus1.5

Output the center point of the position corresponding to the following instruction: instruction annotation. The output should just be the coordinates of a point, in the format [x,y].
(allow refusal)
Output the center point of the position corresponding to the following instruction: {instruction annotation}.
The output should just be the coordinates of a point, in the format [x,y]. Additionally, if the task is infeasible (e.g., the task is not related to the image), the output should be [-1,-1].

Prompt of UI-TARS

Task: instruction annotation

Prompt of UGround

(ground only positive) In the screenshot, where are the pixel coordinates (x, y) of the element corresponding to "instruction annotation"?
(allow negative)
Output the bounding box in the image corresponding to the instruction "{instruction annotation}". If the target does not exist, respond with "Target does not exist".

The MAI-UI-8B gaze settings in Appendix B.8 share the system message below. In each user message, the text is followed by a newline and the screenshot. Braced fields are replaced by the sample's instruction and coordinates, using the raw or normalized rule in Appendix B.8.

Prompt of MAI-UI-8B with Gaze

System message (both settings):

```
You are a GUI grounding agent.
## Task
Given a screenshot and the user's grounding instruction.  Your
task is to accurately locate a UI element based on the user's
instructions.
First, you should carefully examine the screenshot and analyze
the user's instructions, translate the user's instruction into an
effective reasoning process, and then provide the final coordinate.
## Output Format
Return a json object with a reasoning process in
<grounding_think></grounding_think> tags, a [x,y] format coordinate
within <answer></answer> XML tags:
<grounding_think>...</grounding_think>
<answer>
{"coordinate":  [x,y]}
</answer>
## Input instruction
```

User message (raw):

```
I look at [{gx}, {gy}] in this image.  {instruction}
```

User message (normalized):

```
I look at [{norm_x}, {norm_y}] in this image.  {instruction}
```

E.2 Human Performance Evaluation

To verify the validity of our benchmark and the quality of the annotations, three participants who had not been involved in annotation provided informed consent and completed the test on the full dataset. We calculated the average accuracy using the same criteria as those used for model evaluation. The interface used by the participants for the test is shown in Fig. 12.

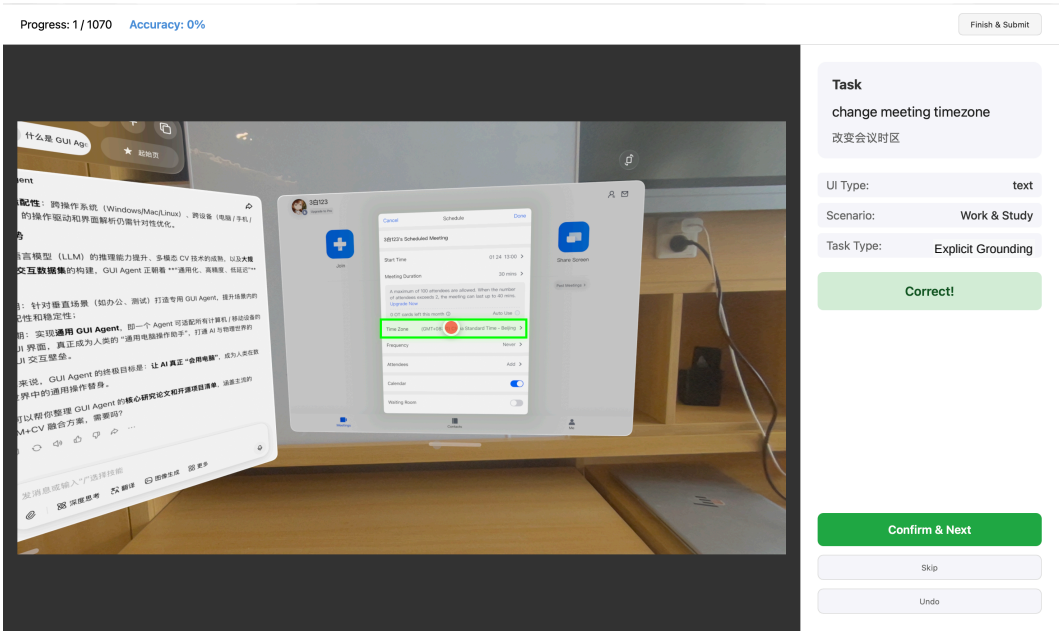


Figure 12: Human performance testing interface on EgoXR-GUI

E.3 Data Synthesis and Fine-tuning For Semantic Grounding

E.3.1 XR-GUI Grounding Data Synthesis Pipeline

To train the semantic reasoning localization model, we design an automated four-stage data synthesis pipeline that transforms egocentric real-world images into an XR virtual-reality fused dataset with semantic localization annotations. The core of the pipeline is Qwen3.6-Plus (Qwen Team, 2026) for scene analysis and instruction generation, supplemented by GPT Image 2 (OpenAI, 2026) for constructing virtual UI screenshots and UI-Venus-1.5-8B (Team et al., 2026) for bounding-box prediction.

Given an egocentric image I, we call the VLM for structured scene analysis, outputting:

- **Scene type** (scene): the environment category, *e.g.*, kitchen, living room, office;
- **User activity** (activity): the action the user is performing, *e.g.*, chopping vegetables, washing dishes;
- **Object list** (objects): major objects in the image, each containing name, color, spatial position in the image (top / bottom / left / right / center), and a textual description.

This information serves as the semantic anchor for subsequent stages.

Inputting the image and scene understanding results to the VLM, we generate K most relevant shopping/service app candidates. Each candidate contains:

- App name (*e.g.*, JD.com, Meituan, Taobao);
- Spatial position of the UI window in the image (*e.g.*, upper-right, lower-right);
- List of related items to display in the app.

These candidates cover diverse shopping scenarios in real XR usage.

Using the image generation model, we synthesize a virtual-reality fused XR screenshot I^{xr} from the egocentric image I, scene information $\mathcal{S}$, and the selected target app app^*. The generation prompt asks the model to preserve the real physical scene while embedding the target app's UI at the specified position with the related items displayed. This step pairs physical-scene content with a task-relevant virtual interface.

Finally, we invoke the VLM on the synthesized XR screenshot I^{xr} to generate two types of localization annotations:

Explicit Instruction (q^{exp}): based on explicit UI element identifiers (*e.g.*, "click the product card labelled `Fresh Okra`"). This type of instruction does not require understanding the physical scene; the model can localize directly.

Semantic Instruction (q^{sem}): based on implicit semantic reasoning from the physical scene (*e.g.*, "view the details of the green vegetable I am currently chopping"). To ensure that semantic instructions truly require physical scene reasoning, we impose two additional constraints:

1. The instruction must not directly contain object names; relational / positional descriptions must be used instead;
2. The correct answer cannot be determined from the UI screenshot alone; the physical scene must be consulted.

We pass each explicit instruction q^{exp} and synthesized screenshot to UI-Venus-1.5-8B to predict the target bounding box. The paired semantic instruction q^{sem} is associated with that same target. Each annotation contains:

1. Target UI element name;
2. Corresponding 2D bounding box $b = (x_1, y_1, x_2, y_2)$, normalized to $[0, 1000]$ click coordinate system;
3. Related physical object (related_object) serving as the semantic anchor.

Algorithm 1 XR-GUI grounding data synthesis pipeline

Require: A set of egocentric images $\mathcal{I}$
Ensure: Annotated dataset $\mathcal{D} = \{(I_i^{xr}, q_i^{exp}, q_i^{sem}, b_i)\}$

Initialize Qwen3.6-Plus client $\mathcal{V}$, GPT Image 2 client $\mathcal{G}$, and UI-Venus-1.5-8B grounder $\mathcal{B}$
Create output directory O
for each image $I \in \mathcal{I}$ **do**
  [Step 1] $\mathcal{S} \leftarrow \mathcal{V}$.analyze_scene$(I)$ {Output: scene description, activity, object list (name, color, spatial position)}
  Save $\mathcal{S}$ to $O/$scene.json
  [Step 2] $\mathcal{C} \leftarrow \mathcal{V}$.suggest_ui_apps$(I, \mathcal{S})$ {Output: K candidate apps $\{(app_j, pos_j, items_j)\}$}
  Save $\mathcal{C}$ to $O/$ui_candidates.json
  Select candidate $app^* \in \mathcal{C}$.apps
  [Step 3] $prompt \leftarrow \mathcal{V}$.build_image_prompt$(I, \mathcal{S}, app^*)$ {Construct XR screenshot generation prompt}
  $I^{xr} \leftarrow \mathcal{G}$.generate_image$(prompt)$ {Generate XR virtual-reality fused screenshot}
  [Step 4] $\mathcal{Q} \leftarrow \mathcal{V}$.generate_instructions$(I^{xr}, \mathcal{S}, app^*)$ {Paired instructions $\{(q^{exp}, q^{sem}, \text{ui_element})\}$}
  $\mathcal{A} \leftarrow \emptyset$
  for each (q^{exp}, q^{sem}, u) in $\mathcal{Q}$ **do**
    $b \leftarrow \mathcal{B}$.ground(I^{xr}, q^{exp})
    $\mathcal{A} \leftarrow \mathcal{A} \cup \{(q^{exp}, q^{sem}, u, b)\}$
  end for
  Save $\mathcal{A}$ to $O/$annotations.json
  Write manifest: $(I, \mathcal{S}, app^*, I^{xr}, \mathcal{A})$
end for
Aggregate all annotations: $\mathcal{D}_{\text{summary}} \leftarrow \bigcup_i \mathcal{A}_i$
Save $\mathcal{D}_{\text{summary}}$ to $O/$summary.json

Table 13: Human audit of 300 synthetic annotations by two reviewers. The two criteria were assessed separately.

Criterion	Passing annotations	Pass rate (%)
Instruction validity	288/300	96
Bounding-box accuracy	288/300	96



Figure 13: **Data synthesis results** This figure illustrates representative samples of our data synthesis pipeline.

Synthetic annotation audit. Two human reviewers audited 300 of the 1,500 synthetic annotations using two criteria. Instruction validity requires the physical scene and UI to support the instruction and identify a valid semantic target. Bounding-box accuracy requires the predicted box to cover the intended GUI element. Both instruction validity and bounding-box accuracy reached 96% on the audited subset (Table 13).

Training and evaluation data. The 500 synthetic XR images and 1,500 semantic annotations are used only for training. We evaluate the base and fine-tuned Qwen3-VL-8B-Instruct models on the real EgoXR-GUI Semantic Grounding set, with the same text/icon categories and point-in-box accuracy metric as the main benchmark. Table 4 reports the corresponding results. The real benchmark screenshots are reserved for evaluation in this experiment.

E.3.2 Training Configuration

We fine-tune Qwen3-VL-8B-Instruct (Bai et al., 2025a) using LoRA (Hu et al., 2022). The model is trained with mixed precision (bfloat16) and Flash Attention 2. We use the AdamW optimizer with cosine learning rate schedule.

Training hyperparameters are summarized in Table 14.

Table 14: Training hyperparameters.

Parameter	Value
Backbone model	Qwen3-VL-8B-Instruct
Training epochs	8
Effective batch size	$2 \times 1 \times 16 = 32$
Per-device batch size	1
Gradient accumulation steps	16
Learning rate	1×10^{-4}
LR scheduler	Cosine (with 50 warmup steps)
Weight decay	0.05
Precision	bfloat16 + Flash Attention 2

E.3.3 LoRA Configuration

We apply LoRA to all four attention projection layers of the LLM backbone. Vision encoder is kept frozen.

Table 15: LoRA parameters.

Parameter	Value
LoRA rank r	128
LoRA alpha	16
LoRA dropout	0.05
Target modules	q_proj, k_proj, v_proj, o_proj
Vision encoder	Frozen (no LoRA)
Gradient checkpointing	Enabled
Trainable params	$\sim$122.7M (1.38% of total)
Total params	8.89B

E.3.4 Image Processing

To balance visual quality and memory efficiency, we limit the maximum number of image patches via the `max_pixels` parameter.

Table 16: Image processing settings.

Parameter	Value
max_pixels	1,003,520 ($\approx 1280 \times 28 \times 28$)
min_pixels	263,424 ($336 \times 28 \times 28$)
Max sequence length	8192

E.3.5 Computational Resources

Fine-tuning is conducted on a single node with 2 NVIDIA RTX 4090 D GPUs (48 GB GDDR6X each), interconnected via PCIe.

- **GPU count**: 2 $\times$ NVIDIA RTX 4090 D (48 GB each)
- **Backend**: NCCL (distributed data parallel)
- **Peak memory per GPU**: $\approx$ 35 GB (model + optimizer + activations)
- **Total training time**: depends on dataset size and epoch count
- **Framework**: PyTorch 2.x + HuggingFace Transformers + PEFT + Flash Attention 2

E.4 Licenses for Existing Assets

The licenses for the assets used in this research are as follows: EgoXR-GUI Benchmark (This Work): We intend to release the EgoXR-GUI dataset and evaluation code under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license upon publication. EPIC-KITCHENS-100 Dataset: The raw video data used for synthetic data generation is licensed under the CC BY-NC 4.0 license. Model Assets: Qwen Series (Qwen2.5-VL, Qwen3-VL): Developed by the Alibaba Qwen team, these models are subject to the official Qwen License Agreement. InternVL 3.5: Released under the MIT License. CogAgent / CogVLM: The code is under the Apache 2.0 license, while the weights follow the CogAgent Model License. OS-Atlas / UGround / OpenCUA: These academic baseline models are typically released under Apache 2.0 or MIT licenses. Open-source Tools and Software: Label Studio: Used for data annotation, licensed under the Apache 2.0 license. PyTorch / Transformers / PEFT: These core libraries are licensed under their respective BSD-style and Apache 2.0 licenses. Apple Developer Tools: Used for data collection, governed by the Apple Developer Program License Agreement.

F Limitations and Future Works

Our work has several limitations. First, although the data collection protocol is designed to reflect realistic XR usage, the final query wording is standardized by annotators from participant-recorded intentions; collecting spontaneous user-issued queries would extend the linguistic diversity of the benchmark. Second, the current benchmark is constructed from static frames instead of continuous video, limiting the ability to evaluate temporal reasoning and interaction dynamics in XR environments. Third, our modeling experiments focus on straightforward supervised fine-tuning, and we do not explore reinforcement learning or other advanced training paradigms that may further improve grounding performance, partially due to the scarcity of the available data.

For future work, we plan to scale XR grounding data generation with synthetic data engines. The current benchmark was collected with eight volunteers on Apple Vision Pro; the selected-case Pico study in Appendix B.7 is a first step toward broader device coverage. Our screenshots capture visible hand–object relations, and Appendix B.8 evaluates gaze coordinates as an additional text input. Temporally aligned video, actions, gaze, and hand trajectories could enable richer grounding tasks and more realistic XR interaction modeling.

G Broader Impact

This work presents EgoXR-GUI, the first benchmark for GUI grounding in XR, addressing a critical gap in the ability of AI agents to reason within hybrid physical-digital environments. By facilitating the development of more intelligent and intuitive XR assistants, this research has the potential to enhance productivity and convenience in both professional and daily activities. Furthermore, improved spatial and semantic grounding can significantly advance accessibility, offering users with cognitive or physical impairments more natural, intent-driven ways to interact with technology. The reliance on egocentric data in XR platforms raises substantial privacy concerns, as it involves capturing personal environments and bystanders during daily activities. Moreover, there are inherent safety risks; if an intelligent agent misinterprets the spatial relationship between virtual elements and the physical world, it could lead to user distraction or incorrect physical actions, potentially resulting in real-world accidents.