

AndroidControl-Curated: Revealing the True Potential of GUI Agents through Benchmark Purification

LEUNG Ho Fai (Kevin)^{†,*}, XI XiaoYan (Sibyl)[†], ZUO Fei (Eric)[†]

[†]BMW ArcherMind Information Technology Co. Ltd. (BA TechWorks)
{kevin.leung,sibyl.xi,eric.zuo}@batechworks.com

Abstract

On-device virtual assistants like Siri and Google Assistant are increasingly pivotal, yet their capabilities are hamstrung by a reliance on rigid, developer-dependent APIs. GUI agents offer a powerful, API-independent alternative, but their adoption is hindered by the perception of poor performance, as even the best models (e.g. Qwen3-VL-235B) scores are capped at around 60% on benchmarks like AndroidControl, far from viability for real-world use.

Our research reveals that issue lies not only with the models but with the benchmarks themselves. We identified notable shortcomings in AndroidControl, including ambiguities and factual errors, which systematically underrates agent capabilities. To address this critical oversight, we enhanced AndroidControl into **AndroidControl-Curated**, a refined version of the benchmark improved through a rigorous purification pipeline. On this enhanced benchmark, state-of-the-art models achieve success rates nearing 75% on complex tasks (15% improvement), reflecting that on-device GUI agents are actually closer to practical deployment than previously thought. We introduce our new SOTA model, **Magma-R1-3B**, post-trained on just 2.4k curated samples using 60 hours of an H20 GPU (approximately \$60). Despite being 200 times smaller in parameters, this model delivers performance comparable to Qwen3-VL-235B. We release both **AndroidControl-Curated** benchmark and **Magma-R1** model to the research community, encouraging adoption of this enhanced benchmark to better reflect model capabilities and accelerate the development of robust, on-device virtual assistants.

Keywords

GUI Agents, Benchmark, Data Quality, Reinforcement Learning, Autonomous Agents

Code and Data: https://github.com/batechworks/AndroidControl_Curated

1 Introduction

Virtual assistants like "Google Assistant" and "Siri" have become a key battleground for mobile phone and automotive manufacturers, increasingly central to the user experience and brand differentiation. Their value is particularly pronounced in cars, where hands-free voice commands for navigation, music, or vehicle functions are critical for safety. This has spurred a shift toward on-device models, such as Google's Gemini Nano (1.8–3.25B parameters) and Apple's 3B parameter foundation model, which are crucial for delivering the fast, private, and offline-capable interactions that users demand.

*Corresponding author. All authors contributed equally to this research.

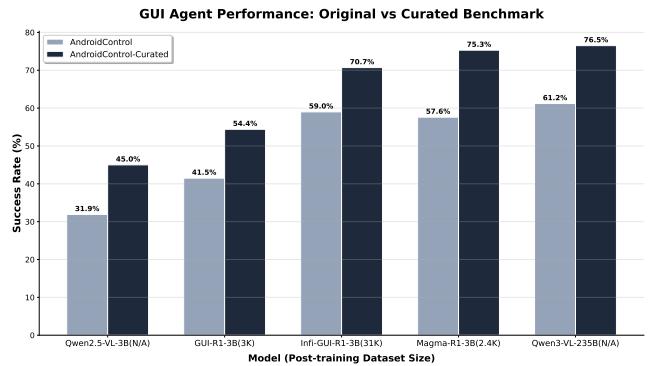


Figure 1: Benchmark Quality Reveals True Model Capabilities. Performance comparison on AndroidControl (original) vs AndroidControl-Curated (purified). Our benchmark purification dramatically improves evaluation accuracy, revealing that compact 3B models (Magma-R1: 75.3%) achieve performance remarkably close to large 235B models (Qwen3-VL-235B: 76.5%) on challenging GUI tasks, up from an upper limit of around 60% of even the world best models. This demonstrates that previous low performance scores were largely due to benchmark deficiencies rather than fundamental model limitations.

However, the capability of these on-device assistants remains severely limited. Due to constrained computational resources, they are restricted to simple tasks like setting timers or launching apps. For more complex, in-app actions, they rely on structured APIs like Apple's App Intents or Google's App Actions. This approach has a fundamental flaw: it is rigid and entirely dependent on developers explicitly providing support, leaving a vast ecosystem of legacy or unsupported applications beyond the assistant's reach.

GUI agents present a far more flexible and powerful alternative [22]. By directly interacting with an application's user interface—reading screen content and programmatically performing actions like tapping and typing—they can automate tasks in any app, regardless of developer support [2, 23]. This paradigm excels at handling the dynamic, multi-step workflows inherent in modern applications, such as filling forms or navigating complex menus, without needing predefined API endpoints that often break with UI updates [6, 32].

Despite this potential, the prevailing view is that on-device GUI agents are not yet viable. Foundation models of a comparable size (3B parameters) have demonstrated low accuracy on high-level

AndroidControl-Curated: A Clean Benchmark for GUI Grounding

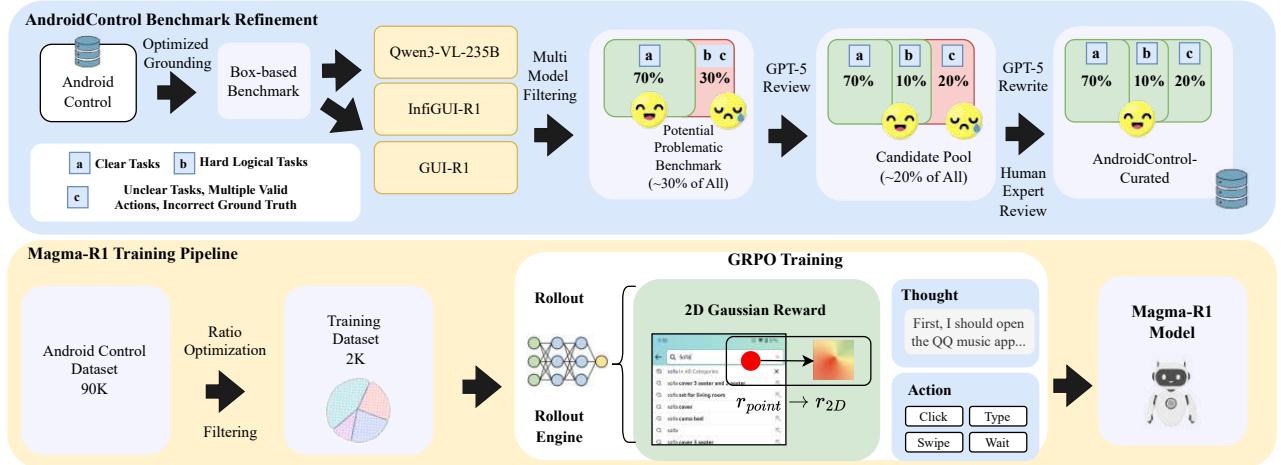


Figure 2: Overview of our integrated pipeline. The upper part shows the Magma-R1 training process with GRPO training, featuring 2D Gaussian reward optimization and ratio optimization for balanced learning. The lower part illustrates the AndroidControl benchmark refinement process, including grounding optimization, multi-model filtering, LLM review and rewrite, leading to the creation of AndroidControl-Curated.

tasks, often hovering in the 60% range on benchmarks like Android-Control even after fine-tuning [9]. Even the latest Qwen3-VL-235B, with 200 times the parameter, seems to reach the same ceiling at around 60% success rate.

This performance is far too low for a reliable user-facing product. This research was born from a direct confrontation with this problem. While developing post-training methodologies for compact vision-language models [19, 26], we encountered perplexing phenomena: training with larger, seemingly higher-quality datasets from the benchmark sometimes led to *reduced* accuracy.

This prompted a deep root-cause analysis, which uncovered a startling truth: the problem lay not with the models, but with the benchmark itself. We found that approximately 30% of the AndroidControl benchmark, though derived from real-world human interactions, was riddled with ambiguities, multiple plausible solutions unaccounted for, and factual inaccuracies. These flaws were systematically penalizing correct and intelligent model behaviors, leading to a significant underestimation of their true capabilities.

To rectify this, we developed a rigorous, semi-automated benchmark purification pipeline, culminating in our primary contribution: **AndroidControl-Curated**. As demonstrated in Figure 1, our purified benchmark provides a more realistic and reliable measure of an agent’s abilities. The results are transformative: on **AndroidControl-Curated**, state-of-the-art model Infi-GUI-R1-3B achieves success rate exceeding 70% (11% increase) on challenging tasks, and the latest Qwen3-VL-235B model even achieves 76.5% success rate (15% increase). This reveals that on-device GUI agents are actually far more capable than previously believed, making them a much nearer-term reality than suggested by earlier benchmarks. This work brings us significantly closer to developing the next generation of GUI-agent-based virtual assistants.

Our main contributions are summarized as follows:

- We identify and systematically analyze critical flaws in a mainstream GUI agent benchmark, revealing that poor benchmark quality is a key bottleneck hindering the perceived viability of on-device GUI agents.
- We propose a reproducible, semi-automated benchmark purification pipeline to address these flaws, providing a methodology for creating more reliable evaluation tools.
- We release **AndroidControl-Curated**, a rigorously purified benchmark. On this benchmark, we show that existing models are far more capable than previously believed, achieving performance that makes on-device GUI agents a practical possibility in the nearer future.
- We open-source **Magma-R1**, a new state-of-the-art model that achieves comparable results to models trained with 13X more data, as shown in Figure 1, demonstrating that benchmark quality is more critical than model scale for GUI agent evaluation.

2 Methodology

Our methodology consists of two main parts. First, we describe the systematic pipeline for constructing our high-quality benchmark, AndroidControl-Curated, designed to more accurately evaluate GUI agents. Second, we detail the training paradigm for our SOTA agent, Magma-R1, which leverages this curated data with our novel reinforcement learning strategies. The entire integrated process is summarized in Algorithm 1 and illustrated in Figure 2.

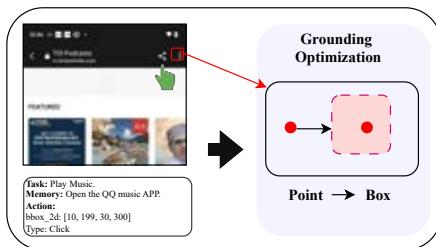


Figure 3: Illustration of our grounding optimization approach. We transform the evaluation from exact point matching to bounding box-based intent alignment. The example shows a GUI task where the agent needs to interact with a UI element, and our method evaluates whether the predicted point falls within the target element’s bounding box rather than matching exact coordinates.

2.1 Systematic Benchmark Purification: The AndroidControl-Curated Pipeline

While existing GUI agent benchmarks like AndroidControl have been invaluable resources for the field [9], we find their evaluation methods and data labels contain systemic flaws. These deficiencies lead to a significant underestimation of agent capabilities and may misdirect future research [1, 27]. To resolve this, we propose a systematic, semi-automated benchmark purification pipeline to create a fairer, more accurate, and more challenging benchmark—**AndroidControl-Curated**. This process involves two core stages: grounding metric refinement and task-level correction.

2.1.1 Stage 1: From Coordinate Matching to Intent Alignment in Grounding Evaluation. **Problem Statement:** Original benchmarks typically evaluate grounding accuracy using **Exact Point Matching** [2, 10]. This approach is fundamentally flawed as it assesses the geometric proximity of predicted and labeled coordinates, not whether the model’s *interaction intent* is correct.

Formally, for a given GUI screenshot I and a ground-truth interaction point $p_{gt} = (x_{gt}, y_{gt})$, the model’s prediction is p_{pred} . The traditional evaluation function, \mathcal{E}_{point} , can be defined as:

$$\mathcal{E}_{point}(p_{pred}, p_{gt}) = \mathbb{I}(\|p_{pred} - p_{gt}\|_2 \leq \tau) \quad (4)$$

where $\mathbb{I}(\cdot)$ is the indicator function and τ is a minimal tolerance threshold, often set to 0. This metric is overly strict, as it ignores the fact that any click within the area of an interactive UI element should be considered a valid execution of a human’s intent. Therefore, \mathcal{E}_{point} fails to accurately measure the model’s understanding of the correct interaction target.

Solution: Bounding-Box-based Intent Alignment. To align the evaluation metric more closely with human interaction logic, we propose a new paradigm based on UI element bounding boxes, as illustrated in Figure 3 [5, 29]. The core idea is to shift the focus of evaluation from the geometric alignment of coordinates to the semantic alignment of intent.

Specifically, we leverage the Document Object Model (DOM) or Accessibility Tree of the GUI [7, 18]. For each interactive UI element

e_k on screen I , we can obtain its minimal bounding box B_k . We define a mapping function $\mathcal{F} : p \rightarrow B$ that maps any point p to the minimal UI element’s bounding box B that contains it.

Consequently, we upgrade the ground-truth point p_{gt} to a more robust **Ground-Truth Interaction Area**, $B_{gt} = \mathcal{F}(p_{gt})$. Our new grounding evaluation function, \mathcal{E}_{bbox} , is defined as:

$$\mathcal{E}_{bbox}(p_{pred}, p_{gt}) = \mathbb{I}(p_{pred} \in B_{gt}) = \mathbb{I}(p_{pred} \in \mathcal{F}(p_{gt})) \quad (5)$$

This function checks if the predicted point p_{pred} falls within the bounding box B_{gt} of the UI element corresponding to the ground-truth point p_{gt} . This refinement offers significant advantages in robustness, intent alignment, and fairness. We refer to the benchmark version with only this refinement applied as **AndroidControl-Curated-Box**.

2.1.2 Stage 2: Task-Level Correction via LLM-Human Collaboration. After addressing the evaluation-level bias, we tackle a deeper challenge: **factual deficiencies** within the benchmark data itself [11, 31]. To eradicate these issues, we designed a three-step, semi-automated task-level correction process.

Step 1: High-Risk Sample Identification via Execution Consensus Failure. To efficiently locate potentially flawed samples, we employ a heuristic strategy based on **Execution Consensus Failure** [4, 28]. The assumption is that if a task cannot be solved by multiple, diverse, high-performing agents—in our case, Qwen3-VL-235B, InfiGUI-R1, and GUI-R1—the task itself is more likely to be flawed than all agents being incapable.

Formally, we define a set of N top-tier GUI agents $\mathcal{M} = \{M_1, \dots, M_N\}$. For any task T_j in the AndroidControl \mathcal{D}_{orig} , we define a success function $S(M, T) \in \{0, 1\}$. The high-risk candidate set \mathcal{D}_{cand} is then identified as:

$$\mathcal{D}_{cand} = \left\{ T_j \in \mathcal{D}_{orig} \mid \sum_{k=1}^N S(M_k, T_j) = 0 \right\} \quad (6)$$

This strategy effectively pinpoints tasks that were failed by every agent in our expert set, significantly narrowing the scope for manual review.

Step 2: Automated Causal Attribution and Correction with LLMs. We leverage a large language model (LLM) as an **Automated Reviewer**, $\mathcal{L}_{reviewer}$ [8, 13]. After establishing a comprehensive **Deficiency Taxonomy** C (e.g., Multiple Valid Actions, Unclear Task, Wrong Ground Truth), we feed each candidate task T_j along with its context and the failed execution traces from our agent set \mathcal{M} to the reviewer. This process is a mapping that yields a structured analysis tuple A_j :

$$A_j = (c_j, \Delta t_j, \Delta g_j, \rho_j) \quad (7)$$

where $c_j \in C$ is the causal attribution of the failure, Δt_j is a proposed revision for the task instruction, Δg_j is a proposed revision for the ground-truth trajectory, and ρ_j is a natural language rationale. This allows for in-depth, scalable analysis and correction proposals.

Step 3: Rigorous Human Expert Verification. Finally, all correction proposals generated by $\mathcal{L}_{reviewer}$ are subject to a **Human-in-the-Loop** verification process [24]. A panel of human experts reviews a large, random sample of the automated corrections, validating

both the causal attribution and the quality of the proposed fixes. This final step ensures the reliability and quality of the resulting benchmark, **AndroidControl-Curated**.

2.2 Training Paradigm of Magma-R1: Optimization via GRPO

To enhance the decision-making and operational precision of GUI agents in complex tasks, we employ a policy optimization-based reinforcement learning framework to train Magma-R1 [15, 16]. We opt for an efficient policy gradient algorithm, which we term Generative REINFORCE with Policy Optimization (GRPO). Its core idea is inspired by Proximal Policy Optimization (PPO), ensuring training stability by constraining the policy update step size [30].

The objective function of GRPO is defined as:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q \sim \mathcal{D}, \{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{G} \sum_{i=1}^G \right. \\ \left. (\min(r_i(\theta)A_i, \text{clip}(r_i(\theta), 1 - \epsilon, 1 + \epsilon)A_i)) \right] \quad (1)$$

where $r_i(\theta) = \frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}$ is the probability ratio for action o_i between the new and old policies, π_θ and $\pi_{\theta_{\text{old}}}$. A_i is the advantage function, which evaluates the quality of action o_i relative to the average for a given query q . ϵ is a hyperparameter that defines the clipping range to prevent excessively large policy updates.

In our implementation, the advantage function A_i is estimated by normalizing the in-batch rewards as follows:

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \dots, r_G\})}{\text{std}(\{r_1, r_2, \dots, r_G\}) + \delta} \quad (2)$$

where r_i is the reward for action o_i , G is the batch size, and δ is a small constant for numerical stability. This method uses the average reward of the batch as a baseline, effectively reducing the variance of the policy gradient.

However, the traditional reinforcement learning paradigm faces two significant challenges in GUI tasks: **Reward Sparsity** and **Data Imbalance**. To address these, we introduce two key optimizations.

2.2.1 Dense Rewards: Gaussian Kernel-based Grounding Reward. In GUI grounding tasks, conventional reward functions (e.g., based on IoU or whether a point is within a bounding box) are typically binary (0 or 1), leading to a severe reward sparsity problem [2, 5]. The model receives positive feedback only for predictions within the exact correct region, while "nearly correct" predictions are treated as complete failures, providing no effective learning signal.

To solve this, we design a smooth reward function based on a 2D Gaussian kernel, which provides a dense and continuous reward signal for grounding tasks [17, 21]. Specifically, for a task requiring the localization of a target point $p_{gt} = (x_{gt}, y_{gt})$, the model predicts a bounding box B_{pred} . We first calculate the center of this box, $c_{pred} = (c_x, c_y)$. The reward function $R_{\text{grounding}}$ is defined as:

$$R_{\text{grounding}}(B_{pred}, p_{gt}) = \exp \left(-\frac{\|c_{pred} - p_{gt}\|_2^2}{2\sigma^2} \right) \quad (3)$$

where $\|\cdot\|_2^2$ denotes the squared Euclidean distance, and σ is a hyperparameter controlling the "tolerance" of the Gaussian function. This function has several advantages:

- **Reward Density:** Any prediction receives a reward in the range (0, 1], decreasing with distance from the ground truth, thus solving the sparsity issue.
- **Smoothness:** The reward function is continuous and differentiable, providing a smooth optimization landscape for the policy gradient.
- **Physical Intuition:** It aligns with human interaction logic—a click slightly off-center but still on the target element is far better than a click that misses entirely.

2.2.2 Balanced Learning: Action Type Proportional Optimization. GUI operation datasets commonly suffer from a severe class imbalance problem, where "click/tap" actions constitute the vast majority, while other critical actions like "type", "scroll", and "swipe" are relatively rare [9, 27]. This long-tailed distribution causes the model to over-focus on high-frequency actions during training, neglecting the learning of low-frequency yet crucial skills.

To promote balanced learning across all action types, we propose an **Action Type Proportional Optimization** strategy [3, 25]. This strategy employs stratified sampling based on a predefined **Target Distribution** when constructing training batches, rather than purely random sampling.

Formally, let $\mathcal{A} = \{a_1, \dots, a_K\}$ be the set of all K action types. In the original dataset $\mathcal{D}_{\text{orig}}$, the empirical distribution is $\mathcal{P}_{\text{orig}} = \{p(a_1), \dots, p(a_K)\}$, where $p(a_{\text{click}})$ is much larger than other probabilities. Our goal is to define a more balanced target distribution $\mathcal{P}_{\text{target}}$, such that:

$$\mathcal{P}_{\text{target}}(a_k) > \mathcal{P}_{\text{orig}}(a_k), \quad \forall a_k \in \{\text{type, scroll, \dots}\}$$

In each training iteration, we sample from data pools stratified by action type, ensuring that the action distribution within each batch \mathcal{B} approximates $\mathcal{P}_{\text{target}}$. This significantly increases the model's exposure to low-frequency actions, leading to a more comprehensive agent proficient in all GUI skills.

3 Experiments

In this section, we present a series of comprehensive experiments to validate the effectiveness of our methodology. Our experiments are designed to answer three core questions: (1) Does our proposed benchmark purification pipeline significantly improve the fairness and accuracy of evaluation? (2) What is the state-of-the-art performance of existing GUI agents on the purified new benchmark? (3) Can our Magma-R1 model, trained on a small amount of high-quality data, outperform baseline models trained on large-scale original data?

3.1 Experimental Setup

Benchmarks: To systematically evaluate the impact of our work, our experiments revolve around a subset of the 'AndroidControl' benchmark [9], introducing two intermediate versions and one final version from our pipeline:

- **AndroidControl:** The original, unmodified 'AndroidControl Hard' benchmark. It serves as the starting point for all our comparisons to reveal the inherent flaws of existing benchmarks.

Algorithm 1 Our Integrated Pipeline: Agent Training and Benchmark Purification

```

1: Input: AndroidControl  $\mathcal{D}_{\text{orig}}$ , un-trained agent parameters  $\theta_0$ 
2: Output: Purified benchmark AndroidControl-Curated, trained agent Magma-R1

3: function CREATEANDROIDCONTROLPRO( $\mathcal{D}_{\text{orig}}$ , expert agents  $\mathcal{M}$ )
4:    $\mathcal{D}_G \leftarrow$  Apply bounding-box evaluation  $\mathcal{E}_{\text{bbox}}$  to  $\mathcal{D}_{\text{orig}}$ 
5:    $\mathcal{D}_{\text{cand}} \leftarrow \{T_j \in \mathcal{D}_G \mid \forall M_k \in \mathcal{M}, S(M_k, T_j) = 0\}$ 
6:    $C_{\text{corrections}} \leftarrow \emptyset$ 
7:   for each  $T_j \in \mathcal{D}_{\text{cand}}$  do
8:      $A_j \leftarrow \mathcal{L}_{\text{reviewer}}(\text{context}(T_j), \text{failure traces})$ 
9:     if  $A_j$  indicates data deficiency then
10:       $C_{\text{corrections}} \leftarrow C_{\text{corrections}} \cup \{A_j\}$ 
11:    end if
12:   end for
13:    $C_{\text{verified}} \leftarrow \text{HumanExpertVerification}(C_{\text{corrections}})$ 
14:   AndroidControl-Curated  $\leftarrow \text{ApplyCorrections}(\mathcal{D}_G, C_{\text{verified}})$ 
15:   return AndroidControl-Curated
16: end function

17: function TRAINMAGMAR1( $\mathcal{D}_{\text{orig}}, \theta_0$ )
18:    $\theta \leftarrow \theta_0$ 
19:   for each training iteration do
20:      $\mathcal{B} \leftarrow \text{StratifiedSample}(\mathcal{D}_{\text{orig}}, \mathcal{P}_{\text{target}})$ 
21:     for each query  $q \in \mathcal{B}$  do
22:       Generate outputs  $\{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot | q)$ 
23:       Calculate dense rewards  $\{r_i\}_{i=1}^G$  using Gaussian kernel (Eq. 3)
24:     end for
25:     Compute advantages  $\{A_i\}_{i=1}^G$  using Eq. 2
26:     Update  $\theta$  by optimizing  $\mathcal{J}_{\text{GRPO}}(\theta)$  (Eq. 1)
27:   end for
28:   return trained agent Magma-R1
29: end function

```

- **AndroidControl-Curated-Box:** This benchmark applies only the **grounding metric refinement** described in Section 3.2.1 to ‘AndroidControl’. The underlying data is identical to the original, but it is evaluated using our more robust $\mathcal{E}_{\text{bbox}}$ standard. This version aims to independently quantify the performance difference caused by the evaluation method itself.
- **AndroidControl-Curated:** This is our final, fully purified benchmark, having undergone the complete pipeline (both grounding metric refinement and task-level correction). It includes both ‘Easy’ and ‘Hard’ subsets from ‘AndroidControl’, providing a more accurate and reliable environment for evaluating GUI agents.

Compared Models: For a thorough comparison, we include several representative GUI agent models as baselines against our proposed model [14, 20]:

- **Infi-GUI-R1:** A public, state-of-the-art model [14] trained on over 31k original ‘AndroidControl’ data points, serving as a key baseline to measure performance improvements.
- **Other Baselines:** We also include several other representative models such as UI-R1 [15], GUI-R1 [16], OS-Atlas, and UI-TARS [20] to ensure the breadth of our evaluation.

► **Part 1: Benchmark Purification (Section 3.1)**

► Stage 1: Grounding refinement
► Stage 2: Task correction

► **Part 2: Agent Training with GRPO (Section 3.2)**

► Proportional sampling

- **Magma-R1 (Ours):** Our proposed model, based on a 3-parameter multimodal model [26] and trained using the GRPO framework described in Section 3.1. Notably, this model was trained on **only 2,400 high-quality samples** selected from AndroidControl-Curated, designed to test the core hypothesis that “data quality trumps quantity.”

Evaluation Metrics: We use two complementary metrics for a comprehensive performance assessment, following established practices in GUI agent evaluation [2, 23]:

- **Success Rate (SR):** The ultimate metric for evaluating an agent’s ability to complete a task [32]. A task is considered successful only if the agent executes all steps correctly and reaches the intended final screen state. This metric measures the agent’s high-level, end-to-end planning and execution capabilities.
- **Grounding Accuracy (GA):** This metric measures the agent’s ability to localize the target for an operation at each step [10]. We consistently use the bounding-box-based intent alignment evaluation function, $\mathcal{E}_{\text{bbox}}$, defined in Section 3.2.1, to calculate GA for all models across all benchmarks. This ensures a fair and consistent evaluation of grounding ability throughout our experiments.

Table 1: Performance comparison of GUI agents on AndroidControl-Curated. Grounding Accuracy (GA) for all models is evaluated using our proposed \mathcal{E}_{bbox} . The best results are in bold, and the second best are underlined. “-” indicates results to be added.

Model	AndroidControl-Curated-Easy			AndroidControl-Curated-Hard		
	Type (%)	Grounding (%)	SR (%)	Type (%)	Grounding (%)	SR (%)
<i>Proprietary Models</i>						
GPT-4o	74.3	0.0	19.4	66.3	0.0	20.8
<i>Open-source Models</i>						
OS-Atlas-4B	91.9	83.8	80.6	84.7	73.8	67.5
UI-R1	62.2	93.6	58.9	54.4	79.3	43.6
GUI-R1-3B	69.5	<u>94.7</u>	67.1	63.1	80.3	54.4
GUI-R1-7B	74.9	95.9	72.7	66.5	82.6	57.5
Infi-GUI-R1 (trained on 31k origin data)	90.2	93.7	<u>87.2</u>	78.5	72.8	70.7
Qwen3-VL-30B	82.8	80.7	70.5	<u>85.9</u>	78.9	70.0
Qwen3-VL-235B	85.1	82.9	74.5	88.2	<u>83.6</u>	76.5
<i>Ours</i>						
Magma-R1	<u>91.3</u>	94.2	88.0	84.2	84.8	<u>75.3</u>

Table 2: Ablation analysis of the benchmark purification process on the Hard subset. SR Impr. (G) shows the SR gain from AndroidControl to AndroidControl-Curated-Box. SR Impr. (T) shows the SR gain from AndroidControl-Curated-Box to the final AndroidControl-Curated. Best results are in bold, second best are underlined.

Model	AndroidControl			AndroidControl-Curated-Box				AndroidControl-Curated			
	Type (%)	Grounding (%)	SR (%)	Type (%)	Grounding (%)	SR (%)	SR Impr. (G)	Type (%)	Grounding (%)	SR (%)	SR Impr. (T)
GUI-R1-3B	57.2	59.0	41.5	59.3	74.0	49.4	+7.9	63.1	80.3	54.4	+5.0
GUI-R1-7B	62.5	65.1	46.3	63.3	76.9	53.2	+6.9	66.5	82.6	57.5	+4.3
Infi-GUI-R1	<u>77.0</u>	57.0	<u>59.0</u>	77.7	69.5	67.6	+8.6	78.5	72.8	70.7	+3.1
Qwen3-VL-235B	67.3	78.3	61.2	82.9	79.9	71.7	+10.5	88.2	<u>83.6</u>	76.5	+4.8
Magma-R1	78.2	58.2	57.6	<u>80.0</u>	<u>77.1</u>	<u>69.1</u>	+11.5	<u>84.2</u>	84.8	<u>75.3</u>	+6.2

3.2 Main Experiment: SOTA Performance on AndroidControl-Curated

We first conduct a comprehensive performance evaluation of all models on the fully purified final benchmark, AndroidControl-Curated. This experiment aims to establish a new standard in two respects: first, to determine the new state-of-the-art performance on the fairest and most accurate GUI benchmark available; and second, to validate our advocated paradigm of "high-quality, small-sample training" through direct comparison with baselines trained on large-scale data [30]. The results are presented in Table 1.

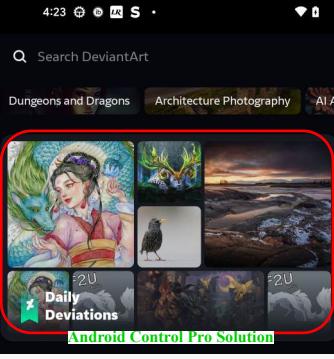
3.3 Ablation Study on Benchmark Purification

To isolate and quantify the specific contribution of each stage in our benchmark purification pipeline, we conduct an ablation study [27]. In this experiment, we evaluate all models on three progressive versions of the ‘AndroidControl-Curated-Hard’ benchmark: ‘AndroidControl’, ‘AndroidControl-Curated-Box’, and the final ‘AndroidControl-Curated’. ‘AndroidControl’ uses the original,

point-based matching for evaluation [2], whereas ‘AndroidControl-Curated-Box’ and ‘AndroidControl-Curated’ both use our intent-aligned, bounding-box-based evaluation (\mathcal{E}_{bbox}). The results, presented in Table 2, clearly illustrate the evolution of model performance as the benchmark is progressively refined.

3.4 Case Study: A Qualitative Analysis of Benchmark Deficiencies

To visually demonstrate the necessity and effectiveness of our task-level correction process, this section provides a qualitative analysis of several typical error cases identified from ‘AndroidControl’ [1]. These cases were first flagged by our automated reviewer, $\mathcal{L}_{reviewer}$, and subsequently confirmed by human experts [24]. Figure 4 illustrates three representative categories of deficiencies we discovered and corrected.



Error Type: Unclear Task

Android Control Problem:

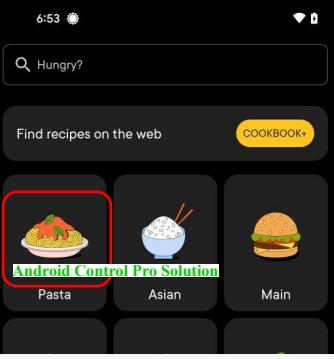
Task: Tosso-dwi is the best literature artist. I get inspiration from the literature artwork created by Tosso-dwi in the DeviantArt app.
 Memory: Step1: Click on the Search section;
 Action Type: Click
 Action Box: [15, 451, 1065, 1096]

X

Android Control Pro Solution:

Task: Explore the Daily Deviations section in the DeviantArt app to find artworks by Tosso-dwi.
 Memory: Step1: Open the DeviantArt app; Step2: Navigate to the homepage.
 Reason: The task description is vague and does not specify any action related to Tosso-dwi's artwork or the search functionality. The memory only mentions clicking the search section, which is not aligned with the ground truth location. The task lacks clarity and does not provide a clear instruction on what needs to be clicked, leading to confusion for the models.

✓



Error Type: Multiple Valid Actions

Android Control Problem:

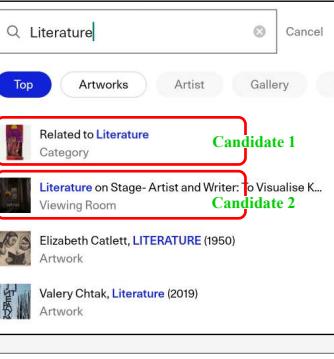
Task: Look for a venison goulash in the kitchen stories app for me, I'm starving.
 Memory: Step1: Click on the Search section;
 Action Type: input_text:venison goulash
 Action Box: None

✓

Android Control Pro Solution:

Action Type: Click
 Action Box: [80, 193, 916, 234]
 Reason: The model predictions focus on clicking the search bar, which is a necessary step before inputting text, suggesting the ground truth should have included this action. Other categories are not applicable because the task and memory are clear, and the model logic is sound.

✓



Error Type: Wrong Ground Truth

Android Control Problem:

Task: Literature art is my favorite art, so look for it on the Artsy app.
 Memory: Step1: Open the Artsy app; Step2: Go to the search section; Step3: Search for Literature;
 Action Type: Swipe: left
 Action Box: None

X

Android Control Pro Solution:

Candidate 1: Action Type: Click Action Box: [80, 193, 916, 234]
 Candidate 2: Action Type: Click Action Box: [80, 256, 916, 330]
 Reason: the ground truth action of swiping left would not accomplish the task objective and appears to be mislabeled. The models that predicted click actions on the search results demonstrate correct understanding of the task flow, while the ground truth fails to provide actionable guidance for completing the stated objective.

✓

Figure 4: Case Study: Three Major Categories of Benchmark Deficiencies. We show representative examples of the systematic errors found in the original AndroidControl benchmark: (Top) Unclear Task - vague task descriptions that lack clear navigation instructions; (Middle) Multiple Valid Actions - multiple valid actions for achieving the same goal; (Bottom) Wrong Ground Truth - incorrect labels that contradict the current UI state. Our AndroidControl-Curated corrections address each category by providing clearer instructions, accepting multiple valid actions, and fixing factual errors.

3.4.1 Case in Point: Three Categories of Systematic Deficiencies. As illustrated in Figure 4, our analysis revealed three primary categories of deficiencies that collectively affect over 70% of the AndroidControl samples, consistent with findings in other evaluation studies [11].

Unclear Task (24.13% of samples): The top panel shows a DeviantArt interface where the task instructs to "get inspiration from literature artwork created by Tosso-dwi," but provides no

clear guidance on how to navigate to this specific content. This type of ambiguous instruction has been identified as a common issue in GUI benchmarks [31]. The ground truth simply indicates clicking the search section, leaving models to guess the subsequent steps. Our correction provides explicit instructions: "Explore the Daily Deviations section in the DeviantArt app to find artworks by Tosso-dwi," giving models a clear, actionable path to success.

Multiple Valid Actions (8.12% of samples): The middle panel demonstrates a cooking app scenario where the task is to "Look for a venison goulash recipe." While the ground truth specifies typing text, clicking on the search bar is an equally valid and necessary prerequisite action. This reflects the multi-modal nature of GUI interactions [29]. Our correction accepts both the search bar click and the subsequent text input as valid actions, recognizing that real user interactions often involve multiple valid pathways to achieve the same goal [7].

Wrong Ground Truth (37.63% of samples): The bottom panel shows the most prevalent issue—factually incorrect labels, a problem that has been noted in other benchmark studies [13]. In an Artsy app interface with literature search results already displayed, the ground truth inexplicably specifies "Swipe: left" with no bounding box, an action that would not advance the stated goal of finding literature art. Our correction identifies that clicking on the displayed search results (such as "Related to Literature") represents the logical next step, replacing the nonsensical ground truth with actionable candidates.

The systematic nature of these deficiencies explains why training on larger datasets sometimes reduced model performance—the benchmark was actively teaching incorrect behaviors [3]. By correcting these issues in AndroidControl-Curated, we observe significant improvements: tasks previously marked as 0% successful due to rigid evaluation criteria now achieve success rates approaching 75%, revealing the true capabilities of existing models and enabling more effective training of new ones [25].

This analysis demonstrates that benchmark quality, not model architecture limitations, was the primary bottleneck preventing the practical deployment of on-device GUI agents [4]. Our systematic correction methodology provides a replicable framework for improving evaluation reliability across the broader field [28].

4 Related Work

4.1 Multimodal Large Language Models

Large Language Models (LLMs) have significantly enhanced the capabilities of AI systems across a wide range of tasks, thanks to their exceptional ability to process complex semantic and contextual information. The remarkable power of LLMs has inspired exploration into their potential for processing multimodal data, particularly images. Multimodal Large Language Models (MLLMs) typically consist of three main components: a pre-trained large language model, a trained modality encoder, and a modality interface that connects the LLM with encoded modality features.

Recent advances have led to the development of specialized models for mobile and GUI understanding. MobileVLM [20] focuses on enhancing both intra- and inter-UI understanding through targeted pre-training tasks, while MobileFlow [19] introduces hybrid visual encoders for variable resolution inputs and multilingual GUI support. These developments have opened new avenues for LLMs in processing GUI tasks across different platforms and languages.

4.2 GUI Agents and Grounding

GUI grounding refers to the task of localizing target elements on a screen based on natural language instructions [2]. Current approaches primarily fall into two paradigms. The first formulates

GUI grounding as a point prediction task, where models directly output coordinates of target elements. Representative works include SeeClick [2], which enhances grounding through pre-training using automated data curation, and ShowUI [10], which introduces UI-guided visual token selection to reduce computational costs.

The second paradigm predicts bounding boxes representing regions that best match instructions, exemplified by works like AriaUI [29], which adopts a pure-vision approach for GUI grounding. Recent advances have also explored universal visual grounding approaches, such as UGround [5], which advocates for human-like embodiment where GUI agents perceive environments entirely visually.

Several representative systems have pioneered the GUI agent area. Mobile-Agent [23] explored mobile app usage through autonomous interactions using visual perception, while more recent works have introduced sophisticated frameworks. UI-TARS [20] presents a native GUI agent model that solely perceives screenshots and performs human-like interactions, incorporating System-2 reasoning and iterative training with reflective online traces.

4.3 Reinforcement Learning for GUI Tasks

Recent advances in GUI agents have shifted toward reinforcement learning-based approaches [17, 21]. Early works introduced binary hit-or-miss rewards for task completion, while more sophisticated approaches have emerged. UI-R1 [15] introduces rule-based action rewards and employs Group Relative Policy Optimization (GRPO) for model optimization. GUI-R1 [16] proposes the first reinforcement learning framework designed to enhance GUI capabilities through unified action space rule modeling.

InfiGUI-R1 [14] advances multimodal GUI agents through an Actor2Reasoner framework, transforming agents from reactive actors to deliberative reasoners using spatial reasoning distillation and reinforcement learning. Other notable approaches include ZeroGUI [28], which automates online GUI learning at zero human cost, and various methods that employ continuous reward mechanisms for fine-grained feedback [30].

Cross-platform compatibility has emerged as a key challenge in GUI agent development [27]. Works like WebVoyager [6] focus on web-based interactions, while Mobile-Agent-E [24] introduces self-evolving mobile assistants with hierarchical multi-agent frameworks. Recent unified approaches attempt to bridge different domains through shared representations [7].

4.4 Benchmark Quality and Data Curation

While substantial progress has been made in developing GUI agents, the quality of evaluation benchmarks has received limited attention [9]. Most existing work focuses on expanding dataset scale or improving model architectures, assuming that existing benchmarks provide reliable evaluation [27]. However, recent studies have begun to highlight the importance of benchmark quality for fair model evaluation.

The issue of benchmark reliability has started to emerge in the GUI domain. Works like LearnAct [11] focus on demonstration-based learning and highlight the importance of high-quality training data, while A3 [1] presents evaluation platforms that address

limitations in existing benchmarks by providing more realistic task scenarios.

Several recent surveys have provided comprehensive overviews of the field. Tang et al. [22] present a systematic examination of (M)LLM-based GUI agents, analyzing their architectural foundations and evaluation methodologies. Liu et al. [12] survey LLM-powered GUI agents specifically in phone automation, highlighting the evolution from script-based automation to intelligent, adaptive systems.

Our work addresses this gap by systematically analyzing and improving the quality of GUI agent benchmarks, demonstrating that benchmark deficiencies can significantly impact the assessment of model capabilities. This aligns with broader efforts in the GUI agent community to improve evaluation practices and benchmark reliability.

5 Conclusion

In this paper, we addressed a fundamental challenge in GUI agent evaluation: systemic deficiencies in widely-used benchmarks that lead to unfair assessment of model capabilities. Our analysis revealed that biased evaluation metrics and factual errors in data labels collectively result in severe underestimation of existing GUI agents' true potential.

We introduced a rigorous, semi-automated benchmark purification pipeline with two key stages: replacing overly strict point-based matching with intent-aligned bounding box evaluation, and employing consensus failure detection, LLM-based analysis, and human expert verification to correct data-level deficiencies. This process produced AndroidControl-Curated, a high-quality benchmark that provides more accurate reflection of model capabilities.

This finding challenges the dominant paradigm and indicates that future research should prioritize enhancing the quality of benchmarks, rather than focusing solely on benchmark performance metrics. Our work also demonstrates that data quality fundamentally trumps data quantity in training high-performance GUI agents. Our Magma-R1 model, trained on only 2,400 carefully curated samples using our proposed GRPO framework, achieved state-of-the-art performance comparable to models trained on over 31k samples of less rigorously curated data.

We release AndroidControl-Curated and Magma-R1 as open-source contributions to catalyze a shift toward more rigorous evaluation practices in GUI agent research. Our work demonstrates that addressing fundamental evaluation issues can unlock the true potential of existing models and enable more reliable progress in autonomous GUI interaction.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback and suggestions. This work was made possible by the generous support of several organizations. We extend our sincere gratitude to Archer-Mind for providing the high-performance computing resources essential for our experiments. We would also like to acknowledge the BMW Group for their significant administrative support. Furthermore, we are grateful to BA Techworks for invaluable technical support and collaboration throughout this project.

References

- [1] Yuxiang Chai, Hanhao Li, Jiayu Zhang, Liang Liu, Guangyi Liu, Guozhi Wang, Shuai Ren, Siyuan Huang, and Hongsheng Li. 2025. A3: Android Agent Arena for Mobile GUI Agents. [doi:10.48550/arXiv.2501.01149](https://doi.org/10.48550/arXiv.2501.01149) arXiv:2501.01149 [cs].
- [2] Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Yantao Li, Jianbing Zhang, and Zhiyong Wu. 2024. SeeClick: Harnessing GUI Grounding for Advanced Visual GUI Agents. [doi:10.48550/arXiv.2401.10935](https://doi.org/10.48550/arXiv.2401.10935) arXiv:2401.10935 [cs] version: 2.
- [3] Gaole Dai, Shiqi Jiang, Ting Cao, Yuanchun Li, Yuqing Yang, Rui Tan, Mo Li, and Lili Qiu. 2025. Advancing Mobile GUI Agents: A Verifier-Driven Approach to Practical Deployment. [doi:10.48550/arXiv.2503.15937](https://doi.org/10.48550/arXiv.2503.15937) arXiv:2503.15937 [cs] version: 2.
- [4] Yue Fan, Handong Zhao, Ruiyi Zhang, Yu Shen, Xin Eric Wang, and Gang Wu. 2025. GUI-Bee: Align GUI Action Grounding to Novel Environments via Autonomous Exploration. [doi:10.48550/arXiv.2501.13896](https://doi.org/10.48550/arXiv.2501.13896) arXiv:2501.13896 [cs].
- [5] Boyu Gou, Ruohan Wang, Boyuan Zheng, Yanan Xie, Cheng Chang, Yiheng Shu, Huan Sun, and Yu Su. 2025. Navigating the Digital World as Humans Do: Universal Visual Grounding for GUI Agents. [doi:10.48550/arXiv.2410.05243](https://doi.org/10.48550/arXiv.2410.05243) arXiv:2410.05243 [cs].
- [6] Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan, and Dong Yu. 2024. WebVoyager: Building an End-to-End Web Agent with Large Multimodal Models. [doi:10.48550/arXiv.2401.13919](https://doi.org/10.48550/arXiv.2401.13919) arXiv:2401.13919 [cs].
- [7] Zhiyuan Huang, Ziming Cheng, Junting Pan, Zhaohui Hou, and Mingjie Zhan. 2025. SpiritSight Agent: Advanced GUI Agent with One Look. [doi:10.48550/arXiv.2503.03196](https://doi.org/10.48550/arXiv.2503.03196) arXiv:2503.03196 [cs] version: 2.
- [8] Wenjia Jiang, Yangyang Zhuang, Chenxi Song, Xu Yang, Joey Tianyi Zhou, and Chi Zhang. 2025. AppAgentX: Evolving GUI Agents as Proficient Smartphone Users. [doi:10.48550/arXiv.2503.02268](https://doi.org/10.48550/arXiv.2503.02268) arXiv:2503.02268 [cs].
- [9] Wei Li, William Bishop, Alice Li, Chris Rawles, Folawayo Campbell-Ajala, Divya Tyamagundlu, and Oriana Riva. 2024. On the Effects of Data Scale on UI Control Agents. [doi:10.48550/arXiv.2406.03679](https://doi.org/10.48550/arXiv.2406.03679) arXiv:2406.03679 [cs].
- [10] Kevin Qinghong Lin, Linjie Li, Difei Gao, Zhengyuan Yang, Shiwei Wu, Zechen Bai, Weixian Lei, Lijuan Wang, and Mike Zheng Shou. 2024. ShowUI: One Vision-Language-Action Model for GUI Visual Agent. [doi:10.48550/arXiv.2411.17465](https://doi.org/10.48550/arXiv.2411.17465) arXiv:2411.17465 [cs] version: 1.
- [11] Guangyi Liu, Pengxiang Zhao, Liang Liu, Zhiming Chen, Yuxiang Chai, Shuai Ren, Hao Wang, Shibo He, and Wenchao Meng. 2025. LearnAct: Few-Shot Mobile GUI Agent with a Unified Demonstration Benchmark. [doi:10.48550/arXiv.2504.13805](https://doi.org/10.48550/arXiv.2504.13805) arXiv:2504.13805 [cs].
- [12] Guangyi Liu, Pengxiang Zhao, Liang Liu, Yaxuan Guo, Han Xiao, Weifeng Lin, Yuxiang Chai, Yu Han, Shuai Ren, Hao Wang, Xiaoyu Liang, Wenhao Wang, Tianze Wu, Linghao Li, Hao Wang, Guanjing Xiong, Yong Liu, and Hongsheng Li. 2025. LLM-Powered GUI Agents in Phone Automation: Surveying Progress and Prospects. [doi:10.48550/arXiv.2504.19838](https://doi.org/10.48550/arXiv.2504.19838) arXiv:2504.19838 [cs].
- [13] Xinyi Liu, Xiaoyi Zhang, Ziyun Zhang, and Yan Lu. [n. d.]. UI-E2I-Synth: Advancing GUI Grounding with Large-Scale Instruction Synthesis. ([n. d.]).
- [14] Yuhang Liu, Pengxiang Li, Congkai Xie, Xavier Hu, Xiaotian Han, Shengyu Zhang, Hongxia Yang, and Fei Wu. 2025. InfiGUI-R1: Advancing Multimodal GUI Agents from Reactive Actors to Deliberative Reasoners. [doi:10.48550/arXiv.2504.14239](https://doi.org/10.48550/arXiv.2504.14239) arXiv:2504.14239 [cs].
- [15] Zhengxi Lu, Yuxiang Chai, Yaxuan Guo, Xi Yin, Liang Liu, Hao Wang, Han Xiao, Shuai Ren, Guanjing Xiong, and Hongsheng Li. 2025. UI-R1: Enhancing Efficient Action Prediction of GUI Agents by Reinforcement Learning. [doi:10.48550/arXiv.2503.21620](https://doi.org/10.48550/arXiv.2503.21620) arXiv:2503.21620 [cs].
- [16] Run Luo, Lu Wang, Wanwei He, and Xiaobo Xia. 2025. GUI-R1 : A Generalist R1-Style Vision-Language Action Model For GUI Agents. [doi:10.48550/arXiv.2504.10458](https://doi.org/10.48550/arXiv.2504.10458) arXiv:2504.10458 [cs].
- [17] Tiange Luo, Lajanugen Logeswaran, Justin Johnson, and Honglak Lee. 2025. Visual Test-time Scaling for GUI Agent Grounding. [doi:10.48550/arXiv.2505.00684](https://doi.org/10.48550/arXiv.2505.00684) arXiv:2505.00684 [cs].
- [18] Shravan Nayak, Xiangru Jian, Kevin Qinghong Lin, Juan A Rodriguez, Montek Kalsi, Rabiuwal Awal, Nicolas Chapados, M Tamer Özsu, Aishwarya Agrawal, David Vazquez, Christopher Pal, Perouz Taslakian, Spandana Gella, and Sai Rajeswar. [n. d.]. UI-Vision: A Desktop-centric GUI Benchmark for Visual Perception and Interaction. ([n. d.]).
- [19] Songqin Nong, Jiali Zhu, Rui Wu, Jiongchao Jin, Shuo Shan, Xiutian Huang, and Wenhao Xu. 2024. MobileFlow: A Multimodal LLM For Mobile GUI Agent. [doi:10.48550/arXiv.2407.04346](https://doi.org/10.48550/arXiv.2407.04346) arXiv:2407.04346 [cs].
- [20] Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, Wanjun Zhong, Kuanye Li, Jiale Yang, Yu Miao, Woyu Lin, Longxiang Liu, Xu Jiang, Qianli Ma, Jingyu Li, Xiaojun Xiao, Kai Cai, Chuang Li, Yaowei Zheng, Chaolin Jin, Chen Li, Xiao Zhou, Minchao Wang, Haoli Chen, Zhaojian Li, Haizhu Yang, Haifeng Liu, Feng Lin, Tao Peng, Xin Liu, and Guang Shi. 2025. UI-TARS: Pioneering Automated GUI Interaction with Native Agents. [doi:10.48550/arXiv.2501.12326](https://doi.org/10.48550/arXiv.2501.12326) arXiv:2501.12326 [cs].

- [21] Fei Tang, Yongliang Shen, Hang Zhang, Siqi Chen, Guiyang Hou, Wenqi Zhang, Wenqiao Zhang, Kaitao Song, Weiming Lu, and Yueteng Zhuang. 2025. Think Twice, Click Once: Enhancing GUI Grounding via Fast and Slow Systems. [doi:10.48550/arXiv.2503.06470](https://doi.org/10.48550/arXiv.2503.06470) arXiv:2503.06470 [cs].
- [22] Fei Tang, Haolei Xu, Hang Zhang, Siqi Chen, Xingyu Wu, Yongliang Shen, Wenqi Zhang, Guiyang Hou, Zeqi Tan, Yuchen Yan, Kaitao Song, Jian Shao, Weiming Lu, Jun Xiao, and Yueteng Zhuang. 2025. A Survey on (M)LLM-Based GUI Agents. [doi:10.48550/ARXIV.2504.13865](https://doi.org/10.48550/ARXIV.2504.13865) Version Number: 2.
- [23] Junyang Wang, Haiyang Xu, Jiabo Ye, Ming Yan, Weizhou Shen, Ji Zhang, Fei Huang, and Jitao Sang. 2024. Mobile-Agent: Autonomous Multi-Modal Mobile Device Agent with Visual Perception. [doi:10.48550/arXiv.2401.16158](https://doi.org/10.48550/arXiv.2401.16158) arXiv:2401.16158 [cs] version: 2.
- [24] Zhenhai long Wang, Haiyang Xu, Junyang Wang, Xi Zhang, Ming Yan, Ji Zhang, Fei Huang, and Heng Ji. 2025. Mobile-Agent-E: Self-Evolving Mobile Assistant for Complex Tasks. [doi:10.48550/arXiv.2501.11733](https://doi.org/10.48550/arXiv.2501.11733) arXiv:2501.11733 [cs].
- [25] Qinzhuo Wu, Wei Liu, Jian Luan, and Bin Wang. 2025. ReachAgent: Enhancing Mobile Agent via Page Reaching and Operation. [doi:10.48550/arXiv.2502.02955](https://doi.org/10.48550/arXiv.2502.02955) arXiv:2502.02955 [cs].
- [26] Qinzhuo Wu, Weikai Xu, Wei Liu, Tao Tan, Liujian Liujianfeng, Ang Li, Jian Luan, Bin Wang, and Shuo Shang. 2024. MobileVLM: A Vision-Language Model for Better Intra- and Inter-UI Understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 10231–10251. [doi:10.18653/v1/2024.findings-emnlp.599](https://doi.org/10.18653/v1/2024.findings-emnlp.599)
- [27] Weikai Xu, Zhizheng Jiang, Yuxuan Liu, Pengzhi Gao, Wei Liu, Jian Luan, Yuanchun Li, Yunxin Liu, Bin Wang, and Bo An. 2025. Mobile-Bench-v2: A More Realistic and Comprehensive Benchmark for VLM-based Mobile Agents. [doi:10.48550/arXiv.2505.11891](https://doi.org/10.48550/arXiv.2505.11891) arXiv:2505.11891 [cs].
- [28] Chenyu Yang, Shiqian Su, Shi Liu, Xuan Dong, Yue Yu, Weijie Su, Xuehui Wang, Zhaoyang Liu, Jinguo Zhu, Hao Li, Wenhui Wang, Yu Qiao, Xizhou Zhu, and Jifeng Dai. 2025. ZeroGUI: Automating Online GUI Learning at Zero Human Cost. [doi:10.48550/arXiv.2505.23762](https://doi.org/10.48550/arXiv.2505.23762) arXiv:2505.23762 [cs].
- [29] Yuhao Yang, Yue Wang, Dongxu Li, Ziyang Luo, Bei Chen, Chao Huang, and Junnan Li. 2024. Aria-UI: Visual Grounding for GUI Instructions. [doi:10.48550/arXiv.2412.16256](https://doi.org/10.48550/arXiv.2412.16256) arXiv:2412.16256 [cs] version: 1.
- [30] Xinbin Yuan, Jian Zhang, Kaixin Li, Zhuoxuan Cai, Lujian Yao, Jie Chen, Enguang Wang, Qibin Hou, Jinwei Chen, Peng-Tao Jiang, and Bo Li. 2025. Enhancing Visual Grounding for GUI Agents via Self-Evolutionary Reinforcement Learning. [doi:10.48550/arXiv.2505.12370](https://doi.org/10.48550/arXiv.2505.12370) arXiv:2505.12370 [cs].
- [31] Bofei Zhang, Zirui Shang, Zhi Gao, Wang Zhang, Rui Xie, Xiaojian Ma, Tao Yuan, Xinxiao Wu, Song-Chun Zhu, and Qing Li. 2025. TongUI: Building Generalized GUI Agents by Learning from Multimodal Web Tutorials. [doi:10.48550/arXiv.2504.12679](https://doi.org/10.48550/arXiv.2504.12679) arXiv:2504.12679 [cs].
- [32] Zhuosheng Zhang and Aston Zhang. 2024. You Only Look at Screens: Multimodal Chain-of-Action Agents. [doi:10.48550/arXiv.2309.11436](https://doi.org/10.48550/arXiv.2309.11436) arXiv:2309.11436 [cs].