

# InfiGUI-G1: Advancing GUI Grounding with Adaptive Exploration Policy Optimization

Yuhang Liu<sup>1,3\*</sup>, Zeyu Liu<sup>2\*</sup>, Shuanghe Zhu<sup>1</sup>, Pengxiang Li<sup>2</sup>, Congkai Xie<sup>3</sup>,  
 Jiasheng Wang<sup>4,3</sup>, Xueyu Hu<sup>1</sup>, Xiaotian Han<sup>5</sup>, Jianbo Yuan<sup>6</sup>, Xinyao Wang<sup>6</sup>,  
 Shengyu Zhang<sup>1†</sup>, Hongxia Yang<sup>2,3†</sup>, Fei Wu<sup>1</sup>

<sup>1</sup>Zhejiang University, Hangzhou, Zhejiang, China

<sup>2</sup>The Hong Kong Polytechnic University, Hong Kong, China

<sup>3</sup>InfiX.ai, Hong Kong, China

<sup>4</sup>The University of Chicago, Chicago, IL, USA

<sup>5</sup>Independent Researcher

<sup>6</sup>Amazon, Seattle, WA, USA

siriusliuyh@gmail.com, sy\_zhang@zju.edu.cn, hongxia.yang@polyu.edu.hk

## Abstract

The emergence of Multimodal Large Language Models (MLLMs) has propelled the development of autonomous agents that operate on Graphical User Interfaces (GUIs) using pure visual input. A fundamental challenge is robustly grounding natural language instructions. This requires a precise *spatial alignment*, which accurately locates the coordinates of each element, and, more critically, a correct *semantic alignment*, which matches the instructions to the functionally appropriate UI element. Although Reinforcement Learning with Verifiable Rewards (RLVR) has proven to be effective at improving *spatial alignment* for these MLLMs, we find that inefficient exploration bottlenecks *semantic alignment*, which prevents models from learning difficult semantic associations. To address this exploration problem, we present Adaptive Exploration Policy Optimization (AEPO), a new policy optimization framework. AEPO employs a multi-answer generation strategy to enforce broader exploration, which is then guided by a theoretically grounded Adaptive Exploration Reward (AER) function derived from first principles of efficiency  $\eta = U/C$ . Our AEPO-trained models, InfiGUI-G1-3B and InfiGUI-G1-7B, establish new state-of-the-art results across multiple challenging GUI grounding benchmarks, achieving significant relative improvements of up to 9.0% against the naive RLVR baseline on benchmarks designed to test generalization and semantic understanding.

**Code** — <https://github.com/InfiXAI/InfiGUI-G1>

**Extended version** — <https://arxiv.org/abs/2508.05731>

## 1 Introduction

The development of autonomous agents capable of operating across the vast landscape of graphical user interfaces (GUIs) is a key frontier in achieving general-purpose

\*These authors contributed equally.

†Corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

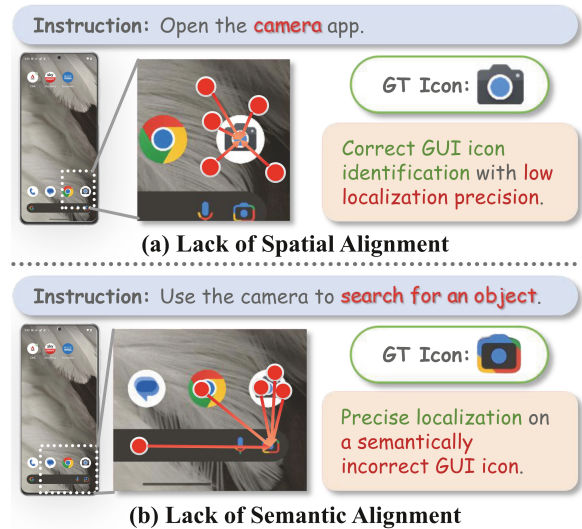


Figure 1: Primary GUI-grounding failure modes. (a) **Spatial-alignment** failure: the model selects the correct icon but localizes it imprecisely. (b) **Semantic-alignment** failure: the model localizes precisely on an incorrect icon due to misinterpreting the instruction. Although RLVR methods have advanced spatial alignment, semantic alignment remains the critical bottleneck for complex GUI tasks—this work is devoted to addressing it.

human-computer interaction (Wang et al. 2024). The success of these agents is fundamentally predicated on a core perceptual task: GUI Grounding. This task involves accurately mapping a natural language instruction to a specific interactive element on a screen. The challenge of GUI Grounding can be deconstructed into two orthogonal dimensions: *Spatial Alignment*, which focuses on the precision of locating an element (i.e., "pointing" accurately), as shown in Fig. 1(a). *Semantic Alignment*, which pertains to the correct-

ness of identifying the appropriate element to interact with (i.e., "pointing" at the right target), as illustrated in Fig. 1(b). Robust and reliable agent performance in complex, real-world scenarios hinges on proficiency in both, with Semantic Alignment being particularly critical.

Current fine-tuning methodologies for multimodal large language models (MLLMs) face major challenges in achieving robust *spatial alignment* and *semantic alignment*. While Supervised Fine-Tuning (SFT) can be effective, it is highly data-intensive and struggles to generalize to unseen UI layouts (Cheng et al. 2024). By contrast, Reinforcement Learning with Verifiable Rewards (RLVR) improves data efficiency by optimizing sequential coordinate generation, which has proven effective at enhancing spatial alignment (Yuan et al. 2025).

However, most of the existing RLVR methods share one limitation: **inefficient exploration**. They rely on the model's current policy to sample actions and thus get stuck on high-confidence errors. This "confidence trap" prevents discovery of low-probability but correct actions, bottlenecking *semantic alignment*. As shown in Fig. 1(b), when the instruction is "Use the camera to search for an object" on a screen displaying various icons, a model with weak semantic understanding may repeatedly select the generic "Camera" button. Standard RLVR would keep sampling this high-confidence but incorrect "Camera" icon, rarely stumbling upon the correct "Google Lens" icon, and thus fail to receive the learning signal necessary to correct its semantic misunderstanding.

We introduce **Adaptive Exploration Policy Optimization (AEPO)**, a novel approach to overcome the exploration bottleneck in standard RL. By integrating the **multi-answer generation strategy**, AEPO drives the model to explore a diverse set of candidate solutions in a single forward pass, addressing the limitations of standard RL, which struggles with low sampling efficiency and the strategy confidence trap. Complemented through the **adaptive exploration reward (AER)**, a non-linear reward signal, AEPO dynamically guides exploration, promoting exploration during failures and convergence upon successes, while avoiding the simplistic or distance-based rewards. Additionally, the **quality-of-exploration penalty** ensures high-quality exploration by penalizing inefficient, near-collinear outputs, fostering true semantic diversity rather than simplistic linear scans in the geometric space. In summary, the key contributions of our work are as follows:

- We present a novel policy-optimization method, **Adaptive Exploration Policy Optimization (AEPO)**, which integrates multi-answer generation into the reinforcement learning framework to boost exploration efficiency for GUI grounding significantly.
- To balance the trade-off between exploration and exploitation, we devise an **Adaptive Exploration Reward (AER)** that incentivizes models to explore both extensively and purposefully.
- Building on the above framework, we introduce the **InfiGUI-G1** series model—3B and 7B variants—whose extensive evaluation across diverse benchmarks establishes a state-of-the-art in the GUI grounding task.

## 2 Related Work

### 2.1 MLLM-based GUI Agents and Grounding

Recently, the paradigm for GUI automation has shifted gradually from brittle, script-based methods to visually driven, human-like approaches (Chen et al. 2025a; Yu et al. 2024). A representative early attempt, OmniParser (Lu et al. 2024), utilizes an MLLM (e.g., GPT-4V (Yang et al. 2023)) to parse visual UI elements in a screenshot into traditional structured data. OS-Atlas (Wu et al. 2024), U-Ground (Gou et al. 2025), and InfiGUIAgent (Liu et al. 2025a) explored hybrid interfaces, intending to achieve robust and flexible performance across diverse environments (Hu et al. 2025; Nguyen et al. 2024). Notably, SeeClick (Cheng et al. 2024) firstly completed GUI tasks via relying solely on screenshots (visual input) and MLLMs, promising greater adaptability and cross-platform universality. However, its approach introduced a new task—GUI grounding—which has been identified as a key metric in this paradigm but also as a primary performance bottleneck.

To address GUI grounding, researchers have advanced a spectrum of techniques that enhance MLLMs' visual-locating capabilities. These include large-scale pre-training on GUI-specific corpora (Qin et al. 2025; Yang et al. 2025a; Wu et al. 2025b), targeted supervised fine-tuning (SFT) (Yang et al. 2025c; Hui et al. 2025), and reasoning-oriented frameworks (Luo et al. 2025; Lee et al. 2025; Wei et al. 2025). In parallel, novel training techniques have been adapted for MLLMs, including coordinate-free methods that generate attention maps instead of explicit coordinates (Wu et al. 2025c), and inference-time optimization strategies that elevate performance without retraining (Wu et al. 2025a), as well as KV-cache-oriented acceleration methods for multimodal models (Jiang et al. 2025a; Li et al. 2025a,b).

### 2.2 Reinforcement Learning in MLLM

Reinforcement learning has rapidly become a potent paradigm for sharpening the reasoning capabilities of multimodal large language models (Liu et al. 2024; Xie et al. 2025). Building on the recent success of DeepSeek-R1 (DeepSeek-AI 2025) in large language models, a succession of vision-centric models, such as Vision-R1 (Huang et al. 2025), Visual-RFT (Liu et al. 2025d), MedVLM-R1 (Pan et al. 2025), InfiMMR (Liu et al. 2025c), demonstrated RL's broad potential across diverse domains (Zhou et al. 2025a; Hu et al. 2024; Jiang et al. 2025b).

In the context of GUI grounding, RL has demonstrated practical applicability through several notable approaches (Liu et al. 2025b; Zhou et al. 2025b; Tang et al. 2025; Lian et al. 2025; Yang et al. 2025b). UI-R1 (Lu et al. 2025) introduces a novel rule-based action reward mechanism that enables model optimization using policy-based algorithms. GUI-R1 (Luo et al. 2025) adopts a unified action space modeling strategy, which extracts and integrates action space categories across different platforms into a cohesive framework. Additionally, self-supervised (Gao, Zhang, and Xu 2025) and self-evolutionary (Yuan et al. 2025) RL methods have been proposed to address the limitations of traditional supervised fine-tuning (SFT), which often relies on large

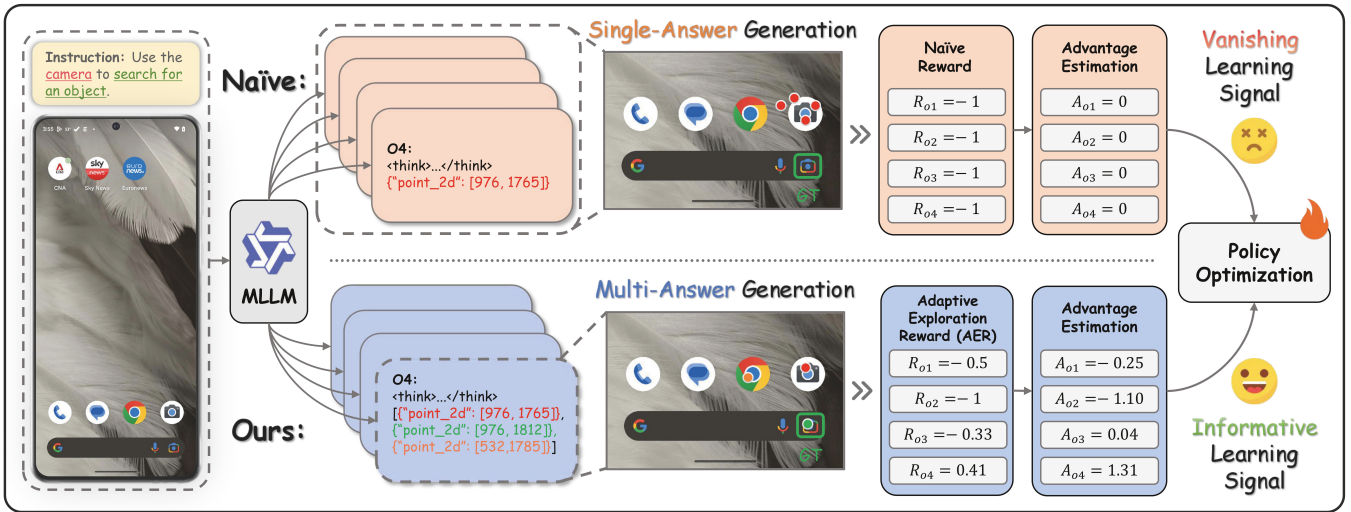


Figure 2: Comparison of AEPO and a naive RL baseline. **Top:** The naive single-answer approach becomes trapped on high-confidence errors, repeatedly sampling the same incorrect action and producing a *vanishing learning signal* when no positive reward is discovered. **Bottom:** AEPO employs multi-answer generation to explore diverse candidates each rollout and an **AER** to derive an *informative learning signal* from their efficiency and correctness. These mechanisms break the exploration bottleneck in GUI agents and enable robust semantic alignment.

amounts of diverse labeled data. Reinforcement fine-tuning (Zhang et al. 2025) also shows promise as a pathway toward integrated training. R-VLM (Park et al. 2025) introduces a two-stage zoom-in grounding process that refines predictions through a zoomed-in view of region proposals. This is combined with an IoU-aware weighted cross-entropy loss to enhance fine-grained perception in grounding tasks. Overall, RL has proven to be an effective and efficient approach for training multi-modal large language models (MLLMs) and advancing GUI grounding performance.

Notably, these methods are constrained by a single-answer generation paradigm, which leads to inefficient exploration and can reinforce the model’s confident but incorrect behaviors. In contrast, our framework employs multi-answer generation to enforce a broader search, which is then guided by our adaptive exploration reward function to provide richer and more effective learning signals.

### 3 Methodology

This section details our proposed AEPO framework. We first formalize the GUI grounding task as a policy optimization problem in §3.1. We then elaborate on the core components of the AEPO framework in §3.2, including multi-answer generation (§3.2), the adaptive exploration reward (§3.2), and the collinear penalty (§3.2). Finally, we present the overall training objective in §3.3.

#### 3.1 Problem Formulation

We formulate GUI grounding as a direct policy optimization problem. The goal is to train a policy  $\pi_\theta$ , represented by an MLLM with parameters  $\theta$ , to generate an action that correctly corresponds to a given context.

- **Context**  $c$ : A tuple  $(\mathcal{S}, \mathcal{I})$ , where  $\mathcal{S}$  is a GUI screenshot and  $\mathcal{I}$  is a natural language instruction.
- **Action**  $a$ : The output generated by the policy, which is a coordinate point  $p = (x, y)$ .
- **Ground Truth**  $B$ : The ground truth bounding box of the target UI element corresponding to the instruction  $\mathcal{I}$ .
- **Policy**  $\pi_\theta(a|c)$ : The policy defines the probability distribution over all possible actions given a context  $c$ .
- **Reward Function**  $R(a, B)$ : A deterministic function that returns a scalar reward. For a generated point  $p$ , the reward is positive if  $p \in B$  and negative otherwise.

The objective is to find the optimal parameters  $\theta^*$  that maximize the expected reward over the data distribution  $\mathcal{D}$ :

$$\theta^* = \arg \max_{\theta} \mathbb{E}_{c \sim \mathcal{D}, a \sim \pi_\theta(\cdot|c)} [R(a, B)] \quad (1)$$

Because the action  $a$  (i.e., the coordinate string) is generated auto-regressively, its sequential generation process is well-suited for optimization with policy gradient algorithms from reinforcement learning, such as Proximal Policy Optimization (PPO, Schulman et al. (2017)), Group Relative Policy Optimization (GRPO, Shao et al. (2024)), or REINFORCE Leave-One-Out (RLOO, Ahmadian et al. (2024)).

#### 3.2 Adaptive Exploration Policy Optimization

To overcome the exploration limitations of the standard formulation, we introduce a novel framework, namely Adaptive Exploration Policy Optimization (AEPO), as depicted in Fig. 2. AEPO enhances the policy optimization process through three synergistic components. The **multi-answer generation** mechanism enhances RL by improving exploration of suboptimal correct answers, overcoming low sampling efficiency and the strategy confidence trap. The **adap-**

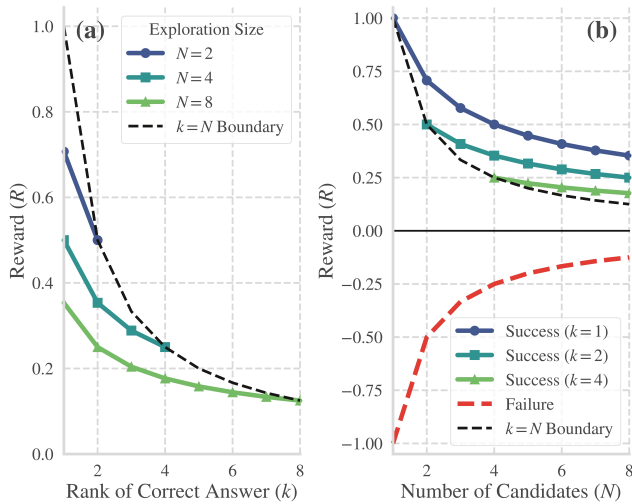


Figure 3: Visualization of the AER function based on the efficiency ratio  $\eta = U/C$ . **(a)** The reward curve increases non-linearly to strongly incentivize selection of the correct answer, i.e., lower rank  $k$ . **(b)** The AER dynamically balances exploration and exploitation: successful trials (green/blue curves) receive higher reward for greater efficiency (smaller candidate set  $N$ ), whereas failures (red curve) incur diminishing penalties to promote broader exploration.

**Adaptive reward function** fosters exploration in response to failure while driving convergence upon success. The **quality-of-exploration penalty** improves exploration quality, ensuring that "multi-answer generation" promotes true diversity in the semantic space, beyond a mere linear scan in the geometric space.

**Multi-Answer Generation.** To fundamentally bypass the exploration bottleneck, our mechanism prompts the model to generate a set of  $N$  candidate points,  $\mathcal{A} = \{p_1, p_2, \dots, p_N\}$ , in a single forward pass (where  $N$  is dynamically determined by the model itself). This forces the model to look beyond its single most confident prediction, significantly increasing the probability of sampling a correct action from the tail of the policy’s distribution, especially for semantically challenging samples.

**Adaptive Exploration Reward.** AER provides an adaptive reward signal to guide the multi-answer exploration process, as visualized in Fig. 3. It is derived from a first-principles model of efficiency,  $\eta = U/C$ , where  $U$  is utility and  $C$  is cost.

- **Utility ( $U$ ):** The utility is defined by the outcome of the exploration. If any point  $p_i \in \mathcal{A}$  falls within the ground truth bounding box  $B$ , the exploration is a success ( $U = +1$ ). Otherwise, it is a failure ( $U = -1$ ), reflecting not only the wasted computational resources but also the risk of guiding the agent into an erroneous state.
- **Cost ( $C$ ):** The cost is modeled as the geometric mean of two components. The **proposal cost**,  $C_p = N$ , represents the effort to generate  $N$  candidates. The **verifica-**

**tion cost**,  $C_v$ , represents the subsequent effort to identify the correct answer. We use the geometric mean,  $C = \sqrt{C_p \cdot C_v}$ , as it appropriately captures the diminishing marginal returns of improving an already high-ranked answer. In case of success,  $C_v = k$  (the rank of the first correct point), leading to  $C_{\text{success}} = \sqrt{N \cdot k}$ . In case of failure, all  $N$  points must be checked, so  $C_v = N$ , and  $C_{\text{failure}} = \sqrt{N \cdot N} = N$ .

This leads to the AER function, which defines the accuracy component of our total reward:

$$R_{\text{accuracy}}(\mathcal{A}, B) = \begin{cases} 1/\sqrt{N \cdot k} & \text{if } \exists p_i \in \mathcal{A} \text{ s.t. } p_i \in B \\ -1/N & \text{otherwise} \end{cases} \quad (2)$$

This reward structure dynamically encourages wider exploration upon failure and rewards efficient, confident predictions upon success.

**Collinear Penalty.** To further improve the quality of exploration, we introduce a penalty for low-quality exploration strategies. If the set of generated points  $\mathcal{A}$  is found to be collinear, we override the accuracy reward with a large negative value,  $R_{\text{accuracy}} = -1$ . Collinearity is determined by checking if the area of the triangle formed by any three points in the set is equal to zero. This discourages the model from adopting trivial, inefficient linear scanning strategies and incentivizes more spatially diverse exploration.

### 3.3 Overall Training Objective

The final reward signal for policy optimization combines a format reward  $R_{\text{format}}$  with the accuracy reward  $R_{\text{accuracy}}$ . The format reward, which is +1 if the output string is correctly structured and 0 otherwise, serves as a prerequisite for any subsequent reward evaluation. The total reward is thus:

$$R_{\text{total}} = R_{\text{format}} + R_{\text{accuracy}} \quad (3)$$

This total reward is then used to compute an advantage estimate,  $\hat{A}$ , which directly guides the update of the policy parameters.

## 4 Experiments

### 4.1 Experimental Setup

**Benchmarks and Metrics.** We evaluate all models on three challenging benchmarks, each chosen to assess distinct capabilities. **MMBench-GUI** (Wang et al. 2025) is a comprehensive benchmark with a hierarchical design of basic and advanced instructions, which we use to evaluate the overall effectiveness of our method across tasks of varying complexity. **ScreenSpot-Pro** (Li et al. 2025c) is a benchmark designed to evaluate performance on high-resolution screens from professional software. Its distinct separation of text-based and icon-based grounding tasks provides a valuable setting to probe a model’s semantic understanding, as icon grounding in particular requires associating abstract symbols with their functions. **UI-Vision** (Nayak et al. 2025) is designed to test generalization across a wide variety of desktop applications, assessing the model’s robustness in diverse, unseen environments. Our primary evaluation metric

(a) Performance on MMBench-GUI														
Model	Windows		MacOS		Linux		iOS		Android		Web		Avg.	$\sigma$
	Basic	Adv.	Basic	Adv.	Basic	Adv.	Basic	Adv.	Basic	Adv.	Basic	Adv.		
GPT-4o (Hurst et al. 2024)	1.5	1.1	8.7	4.3	1.1	1.0	5.1	3.3	2.5	1.4	3.2	2.9	2.9	
Claude-3.7 (Anthropic 2024a)	1.5	0.7	12.5	7.5	1.1	0.0	13.7	10.6	1.4	1.4	3.2	2.3	4.7	
Qwen2.5-VL-7B (Bai et al. 2025)	31.4	16.5	31.3	22.0	21.5	12.2	66.6	55.2	35.1	35.2	40.3	32.5	33.9	
Qwen2.5-VL-72B (Bai et al. 2025)	55.7	33.8	49.9	30.1	40.3	20.9	56.1	28.2	55.6	25.4	68.4	45.8	41.8	
OS-Atlas-Base-7B (Wu et al. 2024)	36.9	18.8	44.4	21.7	31.4	13.3	74.8	48.8	69.6	46.8	61.3	35.4	41.4	
UI-TARS-1.5-7B (Qin et al. 2025)	68.3	39.0	69.0	44.5	64.4	37.8	88.5	69.4	90.5	69.3	81.0	56.5	64.3	
UGround-V1-7B (Gou et al. 2025)	66.8	39.0	71.3	48.6	56.5	31.1	92.7	70.9	93.5	71.0	88.7	64.6	65.7	
InternVL3-72B (Zhu et al. 2025)	70.1	42.6	75.7	52.3	59.2	41.3	93.6	80.6	92.7	78.6	90.7	65.9	72.2	
Naive RLVR-3B	68.6	44.5	78.6	50.0	61.3	39.3	92.4	76.4	91.3	76.1	87.4	63.0	70.9	
Naive RLVR-7B	<u>79.3</u>	<u>58.1</u>	<u>82.3</u>	<u>62.7</u>	64.4	44.9	<u>94.9</u>	<u>89.1</u>	<b>95.5</b>	<u>84.2</u>	<u>92.9</u>	<b>79.5</b>	79.3	
<b>InfiGUI-G1-3B</b>	74.2	47.1	78.8	55.2	65.4	41.8	<b>95.2</b>	78.8	92.1	78.0	89.7	64.3	73.4	0.25
<i>w/ Expl. Success (Avg. N=2.0)</i>	79.7	59.9	86.4	66.8	73.3	54.1	97.1	87.0	96.3	88.7	95.2	75.6	81.6	0.41
<b>InfiGUI-G1-7B</b>	<b>82.7</b>	<b>61.8</b>	<b>83.8</b>	<b>63.9</b>	<b>72.3</b>	<b>52.0</b>	<u>94.9</u>	<b>89.4</b>	<u>95.2</u>	<b>85.6</b>	<b>93.5</b>	<u>76.3</u>	<b>80.8</b>	0.21
<i>w/ Expl. Success (Avg. N=1.6)</i>	87.1	69.1	87.2	76.3	78.5	58.2	98.1	92.4	98.0	91.8	97.1	85.7	86.4	0.11

(b) Performance on ScreenSpot-Pro														
Model	CAD		Dev.		Creative		Scientific		Office		OS		Avg.	$\sigma$
	Text	Icon	Text	Icon	Text	Icon	Text	Icon	Text	Icon	Text	Icon		
GPT-4o (Hurst et al. 2024)	2.0	0.0	1.3	0.0	1.0	0.0	2.1	0.0	1.1	0.0	0.0	0.0	0.8	
Claude Comp. Use (Anthropic 2024b)	14.5	3.7	22.0	3.9	25.9	3.4	33.9	15.8	30.1	16.3	11.0	4.5	17.1	
UI-R1-3B (Lu et al. 2025)	11.2	6.3	22.7	4.1	27.3	3.5	42.4	11.8	32.2	11.3	13.1	4.5	17.8	
ZonUI-3B (Hsieh, Wei, and Yang 2025)	31.9	15.6	24.6	6.2	40.9	7.6	54.8	18.1	57.0	26.4	19.6	7.8	28.7	
GUI-R1-7B (Luo et al. 2025)	23.9	6.3	49.4	4.8	38.9	8.4	55.6	11.8	58.7	26.4	42.1	16.9	31.0	
UI-TARS-7B (Qin et al. 2025)	20.8	9.4	58.4	12.4	50.0	9.1	63.9	<u>31.8</u>	63.3	20.8	30.8	16.9	35.7	
UI-AGILE-7B (Lian et al. 2025)	49.2	14.1	64.3	15.2	53.0	9.8	72.9	25.5	<u>75.1</u>	30.2	45.8	20.2	44.0	
GUI-G <sup>2</sup> -7B (Tang et al. 2025)	<u>55.8</u>	12.5	68.8	17.2	57.1	<u>15.4</u>	<u>77.1</u>	24.5	74.0	32.7	<b>57.9</b>	<u>21.3</u>	47.5	
Naive RLVR-3B	36.0	18.8	63.0	15.2	49.5	13.3	65.3	26.4	64.4	32.1	39.3	16.9	39.8	
Naive RLVR-7B	53.8	17.2	<u>71.4</u>	15.9	<u>60.6</u>	11.9	76.4	26.4	74.6	<u>34.0</u>	54.2	20.2	47.6	
<b>InfiGUI-G1-3B</b>	50.8	<b>25.0</b>	64.9	20.0	51.5	<b>16.8</b>	68.8	<b>32.7</b>	70.6	32.1	49.5	15.7	45.2	0.13
<i>w/ Expl. Success (Avg. N=2.1)</i>	56.9	31.3	70.8	25.5	63.6	23.1	74.3	39.1	79.1	37.7	54.2	19.1	52.0	0.17
<b>InfiGUI-G1-7B</b>	<b>57.4</b>	<u>23.4</u>	<b>74.7</b>	<b>24.1</b>	<b>64.6</b>	<u>15.4</u>	<b>80.6</b>	<u>31.8</u>	<b>75.7</b>	<b>39.6</b>	<u>57.0</u>	<b>29.2</b>	<b>51.9</b>	0.48
<i>w/ Expl. Success (Avg. N=2.0)</i>	65.5	26.6	85.1	30.3	71.2	20.3	84.7	33.6	81.4	47.2	60.7	37.1	58.0	0.24

(c) Performance on UI-Vision											
Model	Grouped by Category						Grouped by Setting			Overall	$\sigma$
	Edu.	Browser	Dev.	Prod.	Creative	Entert.	Basic	Func.	Spatial		
GPT-4o (Hurst et al. 2024)	1.5	0.0	2.2	1.1	0.8	4.2	1.6	1.5	1.0	1.4	
Claude-3.7-Sonnet (Anthropic 2024a)	6.1	9.8	8.0	9.4	7.7	8.3	9.5	7.7	7.6	8.3	
Qwen2.5-VL-7B (Bai et al. 2025)	0.5	0.0	1.2	0.9	0.5	1.0	1.2	0.8	0.5	0.9	
UGround-v1-7B (Gou et al. 2025)	10.4	28.7	17.5	12.2	8.6	18.2	15.4	17.1	6.3	12.9	
UGround-v1-72B (Gou et al. 2025)	22.4	35.7	27.6	21.6	<b>18.3</b>	38.0	27.9	26.7	<b>14.9</b>	23.2	
Aguvis-7B (Xu et al. 2025)	13.1	30.8	17.1	12.1	9.6	24.0	17.8	18.3	5.1	13.7	
UI-TARS-7B (Qin et al. 2025)	14.2	35.0	19.7	18.3	11.1	38.5	20.1	24.3	8.4	17.6	
UI-TARS-72B (Qin et al. 2025)	<u>24.8</u>	40.5	<u>27.9</u>	<b>26.8</b>	<u>17.8</u>	41.1	31.4	30.5	<u>14.7</u>	<u>25.5</u>	
Naive RLVR-3B	18.5	37.8	21.8	19.6	12.8	42.7	27.4	24.6	7.3	19.4	
Naive RLVR-7B	23.5	42.7	27.4	24.5	16.2	<u>50.5</u>	<u>32.9</u>	<u>30.7</u>	10.1	24.1	
<b>InfiGUI-G1-3B</b>	22.6	<u>43.4</u>	24.3	22.6	14.0	47.4	31.2	28.0	8.2	22.0	0.20
<i>w/ Expl. Success (Avg. N=2.1)</i>	29.3	<u>51.7</u>	30.5	31.7	20.5	59.9	39.2	36.7	14.6	29.7	0.29
<b>InfiGUI-G1-7B</b>	<b>25.5</b>	<b>46.2</b>	<b>29.6</b>	<u>26.7</u>	17.6	<b>52.1</b>	<b>36.2</b>	<b>31.9</b>	11.5	<b>26.1</b>	0.05
<i>w/ Expl. Success (Avg. N=2.1)</i>	35.4	52.4	35.5	37.3	23.3	66.1	44.4	40.7	19.5	34.4	0.12

Table 1: Main performance comparison on (a) MMBench-GUI, (b) ScreenSpot-Pro, and (c) UI-Vision. We report Top-1 accuracy (%); for InfiGUI-G1 models, only the first generated answer is evaluated. Best and second-best results are shown in **bold** and underlined, respectively. For our models, we also report the *Exploration Success Rate* with the average number of generated candidates (Avg.  $N$ ), and standard deviation  $\sigma$  over 5 runs.

is **Accuracy**, where a prediction is considered correct if its coordinate point falls within the ground truth bounding box. For methods that output a bounding box, its center point is used. To demonstrate the high success rate of our exploration strategy, we also report the **Exploration Success Rate** for our InfiGUI-G1 models, where a sample is marked as a success if at least one of the generated candidate points is correct.

**Baselines.** To ensure a fair and rigorous comparison, we establish two sets of baselines. First, for controlled analysis, we train Naive RLVR models for both model sizes as **internal baselines**. These baselines are trained using the exact same dataset and optimized hyperparameters as our core models. Second, to position our work within the broader literature, we compare it against several state-of-the-art models from recent works.

**Implementation Details.** Our InfiGUI-G1 models are built upon the open-source **Qwen2.5-VL-3B-Instruct** and **Qwen2.5-VL-7B-Instruct** as backbones. For the RLVR training phase, we adopt the RLOO algorithm (Ahmadian et al. 2024), which effectively reduces the variance of policy gradient estimates by employing the average reward of other samples within the same batch as a baseline. This “leave-one-out” strategy obviates the need for training a separate critic model. The RLOO policy gradient  $\nabla_{\theta} J(\theta)$  is estimated as:

$$\nabla_{\theta} J(\theta) \approx \frac{1}{n} \sum_{i=1}^n \left[ R(y_{(i)}, x) - \frac{1}{n-1} \sum_{j \neq i} R(y_{(j)}, x) \right] \cdot \nabla_{\theta} \log \pi_{\theta}(y_{(i)} | x)$$

where  $n$  is the number of output sequences  $y_{(i)}$  sampled from the policy  $\pi_{\theta}$  given input  $x$ . Across all experiments, we employ a reasoning prompting paradigm, instructing the model to generate its reasoning process within `<think>` `</think>` tags before providing the final answer.

**Training Details.** Our training data is a mixture sampled from several public GUI datasets, including Widget Caption (Li et al. 2020b), OmniAct (Kapoor et al. 2024), GUICourse (Chen et al. 2025b), OS-ATLAS (Wu et al. 2024), ShowUI (Lin et al. 2025), and RICO RCA (Li et al. 2020a), resulting in approximately 44K samples. Following common practices in RLVR to focus training on more challenging instances, we apply a data filtering strategy: for each sample, we generate 8 responses with a temperature of 1.0; if all 8 are correct, the sample is deemed too easy and is excluded. All models were trained on 16 H800 GPUs. Key training parameters include a learning rate of  $1 \times 10^{-6}$ , a rollout batch size of 128, and an RLOO rollout number of  $n = 8$ . We train for 3 epochs.

## 4.2 Main Results

We present the main results of our evaluation in Table 1. The results consistently show that our InfiGUI-G1 models establish new state-of-the-art performance among open-source models in both the 3B and 7B parameter categories. Notably, our models also exhibit competitive or superior performance against several proprietary models with significantly larger

Model Configuration	Acc.	Expl. Succ.	# Answers
<i>3B Models</i>			
<b>InfiGUI-G1 (Full Model)</b>	<b>45.2</b>	<b>52.0</b>	2.1
w/o Multi-Answer (Naive)	39.8	-	1.0
w/o AER (use naive reward)	38.4	42.1	1.9
w/o AER’s rank factor $k$	38.1	47.6	2.5
w/o Collinear Penalty	35.3	44.1	6.6
<i>7B Models</i>			
<b>InfiGUI-G1 (Full Model)</b>	<b>51.9</b>	<b>58.0</b>	2.0
w/o Multi-Answer (Naive)	47.6	-	1.0
w/o AER (use naive reward)	41.4	45.5	1.9
w/o AER’s rank factor $k$	44.0	50.5	1.9
w/o Collinear Penalty	37.0	43.8	8.2

Table 2: Ablation study on the **ScreenSpot-Pro** benchmark. We compare model variants by Top-1 Accuracy (%), *Exploration Success Rate* (%), and average number of answers per sample. Best results within each group are shown in **bold**.

parameter counts, highlighting the efficacy and efficiency of our proposed AEPO framework.

The comparison with our internal baselines reveals that InfiGUI-G1 consistently and substantially outperforms the Naive RLVR model across all benchmarks. This direct comparison suggests that the performance gains can be attributed to the architectural and methodological improvements introduced by AEPO. Furthermore, our models demonstrate strong performance against other SOTA methods, including those based on SFT (e.g., UGround, OS-Atlas), many of which require training data exceeding 1M samples. In contrast, our approach achieves these competitive results using 44K instances, underscoring its data efficiency. Our results also show strong performance against other RLVR approaches that utilize IoU or distance-based rewards (e.g., GUI-R1, GUI-G<sup>2</sup>).

Our method demonstrates strong generalization capabilities by achieving consistently high performance across multiple benchmarks with distinct focuses (e.g., UI-Vision, ScreenSpot-Pro). Crucially, these benchmarks contain many applications and scenarios not present in our training data, indicating that AEPO fosters a robust understanding rather than overfitting. The benefits of AEPO in enhancing semantic understanding appear particularly pronounced on the ScreenSpot-Pro benchmark. Here, our models show a more substantial improvement on icon-based grounding tasks than on text-based ones when compared to the Naive RLVR baseline, suggesting that AEPO’s enhanced exploration is especially beneficial for tasks requiring association of abstract visual symbols with their functions.

To assess statistical significance, we conducted McNemar’s test comparing InfiGUI-G1 and the Naive RLVR baseline. Across all three benchmarks, all p-values are below  $1e-3$ , confirming that the improvements are statistically significant.

Method	3B Models	7B Models
Naive RLVR (pass@2)	41.7	49.8
Naive RLVR (pass@4)	43.5	52.1
<b>InfiGUI-G1 (Expl. Succ.)</b>	<b>52.0</b>	<b>58.0</b>
↪ Avg. $N$	2.1	2.0

Table 3: Exploration efficiency (%) on ScreenSpot-Pro. Our single-pass success rate surpasses the baseline’s multi-pass rate.

### 4.3 Ablation Studies

To dissect the contribution of each component within our AEPO framework, we conduct a series of ablation studies on the ScreenSpot-Pro benchmark. As its icon-based grounding tasks directly probe semantic understanding, this benchmark provides a clear setting to evaluate our design choices. The results are summarized in Table 2.

The results reveal a clear logic behind AEPO’s design. Removing **multi-answer generation** (‘w/o Multi-Answer’) leads to a significant performance drop, confirming that enabling exploration is the necessary first step. However, this exploration must be guided effectively, as replacing our **AER** with a naive reward (‘w/o AER’) causes a further decline. The importance of AER’s ranking factor  $k$  is particularly insightful; removing it (‘w/o  $k$ ’) results in a model that often finds the correct answer (high Expl. Succ.) but fails to rank it first (low Acc.), demonstrating that  $k$  is crucial for teaching the model **confidence** in its correct discoveries. Finally, the **collinear penalty** proves essential for ensuring the **quality** of exploration. Without it, the model adopts a degenerate strategy of generating numerous low-quality answers (high # of answers) while accuracy plummets, showing the penalty is critical for preventing reward hacking.

### 4.4 Analysis of AEPO’s Effectiveness

**Adaptive Exploration Strategy.** We investigate if the model learns an adaptive exploration strategy. A clear correlation emerges between benchmark difficulty (indicated by model accuracy) and exploratory behavior. Our 7B model generates the most answers on the hardest benchmark (UI-Vision: 26.1% Acc, 2.1 answers) and the fewest on the easiest (MMBench-GUI: 80.8% Acc, 1.6 answers). This suggests AEPO learns to adaptively allocate exploratory resources based on task complexity.

To further understand the mechanisms of AEPO, we conduct three targeted analyses.

**Exploration Efficiency.** We then evaluate the quality and efficiency of AEPO’s exploration. Our InfiGUI-G1 models on ScreenSpot-Pro generate approximately two candidate answers per instance on average. To contextualize this, we compare our single-pass Exploration Success Rate against the multi-pass ‘pass@ $k$ ’ accuracy of the Naive RLVR baseline. As detailed in Table 3, the results are compelling. Even when the Naive RLVR model is allowed four independent attempts (‘pass@4’), its success rate in finding a correct answer is still significantly lower than that of our InfiGUI-G1,

Difficulty Subset	3B Models		7B Models	
	Naive RLVR	InfiGUI-G1 (Ours)	Naive RLVR	InfiGUI-G1 (Ours)
Easy	100	<b>100</b>	100	<b>100</b>
Middle	75.9	<b>78.9 (+4.0%)</b>	72.6	<b>78.4 (+8.0%)</b>
Hard	25.5	<b>31.4 (+23.1%)</b>	10.8	<b>17.4 (+61.1%)</b>

Table 4: Accuracy (%) on ScreenSpot-Pro subsets of varying difficulty. AEPO’s advantage is most significant on ‘hard’ samples.

which achieves a higher success rate in a single pass with only about two attempts. This demonstrates that AEPO’s multi-answer generation is not merely about increasing the number of tries, but about performing a more structured and efficient exploration of the action space.

**Performance on Hard-to-Explore Samples.** Finally, to validate our core hypothesis that AEPO resolves the exploration bottleneck, we designed an experiment to analyze performance on samples of varying difficulty. We partitioned the ScreenSpot-Pro test set by first using the base MLLM to generate 16 stochastic responses for each sample. Samples were then labeled as ‘hard’ if the base model failed all 16 times, ‘easy’ if it succeeded every time, and ‘middle’ otherwise. The ‘hard’ subset therefore represents samples that are highly unlikely to be answered correctly through naive exploration. As shown in Table 4, we then compared InfiGUI-G1 against the Naive RLVR baseline on these subsets. While our model improves performance across the board, the most significant gains are concentrated on the ‘hard’ subset. On these critical samples, our 7B model achieves a relative improvement of over 60%. This provides direct evidence that AEPO effectively creates learning signals for previously “unlearnable” samples, addressing the fundamental limitation we set out to solve.

## 5 Conclusion

In this work, we addressed the critical challenge of enhancing semantic alignment in MLLM-based GUI agents, identifying the inefficient exploration of standard RLVR as a key bottleneck. We proposed AEPO, a policy optimization framework that integrates multi-answer generation with a theoretically-grounded AER function to enable effective exploration. Our model, InfiGUI-G1, achieves state-of-the-art performance, and our comprehensive analyses confirm that its effectiveness stems from its ability to adapt its exploration strategy, its high efficiency compared to naive sampling, and its success in creating learning signals for previously “unlearnable” samples.

Limitations of our work include the computational overhead from multi-answer generation and a performance ceiling imposed by the backbone MLLM’s visual capabilities, which could be addressed in future work by exploring more efficient sampling strategies and integration with more advanced visual encoders.

## Acknowledgments

Work by Yuhang Liu and Jiasheng Wang was done during an internship at InfiX.ai. Work by Jianbo Yuan and Xinyao Wang was done outside of Amazon. This work was supported by the National Natural Science Foundation of China (No. 62402429, U24A20326, 62441236), the Key Research and Development Program of Zhejiang Province (No. 2025C01026), the Ningbo Yongjiang Talent Introduction Programme (2023A-397-G), the Young Elite Scientists Sponsorship Program by CAST (2024QNR001), and the Zhejiang University Education Foundation Qizhen Scholar Foundation.

## References

- Ahmadian, A.; Cremer, C.; Gallé, M.; Fadaee, M.; Kreutzer, J.; Pietquin, O.; Üstün, A.; and Hooker, S. 2024. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *arXiv preprint arXiv:2402.14740*.
- Anthropic. 2024a. Claude 3.7 Sonnet System Card. <https://assets.anthropic.com/m/785e231869ea8b3b/original/claude-3-7-sonnet-system-card.pdf>. Accessed: 2025-08-02.
- Anthropic. 2024b. Developing a computer use model. <https://www.anthropic.com/news/developing-computer-use>. Accessed: 2025-04-12.
- Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Chen, J.; Hu, J.; Wang, G.; Jiang, Z.; Zhou, T.; Chen, Z.; and Lv, C. 2025a. TaoAvatar: Real-Time Lifelike Full-Body Talking Avatars for Augmented Reality via 3D Gaussian Splatting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, 10723–10734. Computer Vision Foundation / IEEE.
- Chen, W.; Cui, J.; Hu, J.; Qin, Y.; Fang, J.; Zhao, Y.; Wang, C.; Liu, J.; Chen, G.; Huo, Y.; et al. 2025b. GUICourse: From General Vision Language Model to Versatile GUI Agent. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 21936–21959.
- Cheng, K.; Sun, Q.; Chu, Y.; Xu, F.; Li, Y.; Zhang, J.; and Wu, Z. 2024. SeeClick: Harnessing gui grounding for advanced visual gui agents. *arXiv preprint arXiv:2401.10935*.
- DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv:2501.12948*.
- Gao, L.; Zhang, L.; and Xu, M. 2025. UIShift: Enhancing VLM-based GUI Agents through Self-supervised Reinforcement Learning. *arXiv:2505.12493*.
- Gou, B.; Wang, R.; Zheng, B.; Xie, Y.; Chang, C.; Shu, Y.; Sun, H.; and Su, Y. 2025. Navigating the Digital World as Humans Do: Universal Visual Grounding for GUI Agents. *arXiv:2410.05243*.
- Hsieh, Z.; Wei, T.-J.; and Yang, S. 2025. ZonUI-3B: A Lightweight Vision-Language Model for Cross-Resolution GUI Grounding. *arXiv:2506.23491*.
- Hu, X.; Xiong, T.; Yi, B.; Wei, Z.; Xiao, R.; Chen, Y.; Ye, J.; Tao, M.; Zhou, X.; Zhao, Z.; et al. 2025. Os agents: A survey on mllm-based agents for general computing devices use. *arXiv preprint arXiv:2508.04482*.
- Hu, X.; Zhao, Z.; Wei, S.; Chai, Z.; Ma, Q.; Wang, G.; Wang, X.; Su, J.; Xu, J.; Zhu, M.; et al. 2024. InfiAgent-DABench: evaluating agents on data analysis tasks. In *Proceedings of the 41st International Conference on Machine Learning, 19544–19572*.
- Huang, W.; Jia, B.; Zhai, Z.; Cao, S.; Ye, Z.; Zhao, F.; Xu, Z.; Hu, Y.; and Lin, S. 2025. Vision-R1: Incentivizing Reasoning Capability in Multimodal Large Language Models. *arXiv:2503.06749*.
- Hui, Z.; Li, Y.; Zhao, D.; Chen, T.; Banbury, C.; and Koishida, K. 2025. WinClick: GUI Grounding with Multimodal Large Language Models. *arXiv:2503.04730*.
- Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Jiang, Z.; Li, K.; Zhou, Y.; Liu, S.; Wang, Z.; Lv, C.; and Zhang, S. 2025a. PureKV: Plug-and-Play KV Cache Optimization with Spatial-Temporal Sparse Attention for Vision-Language Large Models. *arXiv:2510.25600*.
- Jiang, Z.; Xu, J.; Zhang, S.; Shen, T.; Li, J.; Kuang, K.; Cai, H.; and Wu, F. 2025b. FedCFA: Alleviating Simpson’s Paradox in Model Aggregation with Counterfactual Federated Learning. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA, 17662–17670*.
- Kapoor, R.; Butala, Y. P.; Russak, M.; Koh, J. Y.; Kamble, K.; Al-Shikh, W.; and Salakhutdinov, R. 2024. Omniact: A dataset and benchmark for enabling multimodal generalist autonomous agents for desktop and web. In *European Conference on Computer Vision*, 161–178. Springer.
- Lee, H.; Kim, J.; Kim, B.; Tack, J.; Jo, C.; Lee, J.; Park, C.; In, S.; Shin, J.; and Yoo, K. M. 2025. ReGUIDE: Data Efficient GUI Grounding via Spatial Reasoning and Search. *arXiv:2505.15259*.
- Li, K.; Jiang, Z.; Shen, Z.; Wang, Z.; Lv, C.; Zhang, S.; Wu, F.; and Wu, F. 2025a. MadaKV: Adaptive Modality-Perception KV Cache Eviction for Efficient Multimodal Long-Context Inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, 13306–13318. Association for Computational Linguistics.
- Li, K.; Xiong, Y.; Jiang, Z.; Zhou, Y.; Wang, Z.; Lv, C.; and Zhang, S. 2025b. FlowMM: Cross-Modal Information Flow Guided KV Cache Merging for Efficient Multimodal Context Inference. *arXiv:2511.05534*.
- Li, K.; Ziyang, M.; Lin, H.; Luo, Z.; Tian, Y.; Ma, J.; Huang, Z.; and Chua, T.-S. 2025c. ScreenSpot-Pro: GUI Grounding for Professional High-Resolution Computer Use. In *Workshop on Reasoning and Planning for Large Language Models*.
- Li, Y.; He, J.; Zhou, X.; Zhang, Y.; and Baldrige, J. 2020a. Mapping Natural Language Instructions to Mobile UI Action Sequences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8198–8210.
- Li, Y.; Li, G.; He, L.; Zheng, J.; Li, H.; and Guan, Z. 2020b. Widget Captioning: Generating Natural Language Description for Mobile User Interface Elements. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 5495–5510.
- Lian, S.; Wu, Y.; Ma, J.; Song, Z.; Chen, B.; Zheng, X.; and Li, H. 2025. UI-AGILE: Advancing GUI Agents with Effective Reinforcement Learning and Precise Inference-Time Grounding. *arXiv preprint arXiv:2507.22025*.
- Lin, K. Q.; Li, L.; Gao, D.; Yang, Z.; Wu, S.; Bai, Z.; Lei, S. W.; Wang, L.; and Shou, M. Z. 2025. Showui: One vision-language-action model for gui visual agent. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 19498–19508.
- Liu, Y.; Hu, X.; Zhang, S.; Chen, J.; Wu, F.; and Wu, F. 2024. Fine-Grained Guidance for Retrievers: Leveraging LLMs’

Feedback in Retrieval-Augmented Generation. *arXiv preprint arXiv:2411.03957*.

Liu, Y.; Li, P.; Wei, Z.; Xie, C.; Hu, X.; Xu, X.; Zhang, S.; Han, X.; Yang, H.; and Wu, F. 2025a. InfiGUIAgent: A Multimodal Generalist GUI Agent with Native Reasoning and Reflection. In *ICML 2025 Workshop on Computer Use Agents*.

Liu, Y.; Li, P.; Xie, C.; Hu, X.; Han, X.; Zhang, S.; Yang, H.; and Wu, F. 2025b. Infigui-r1: Advancing multimodal gui agents from reactive actors to deliberative reasoners. *arXiv preprint arXiv:2504.14239*.

Liu, Z.; Liu, Y.; Zhu, G.; Xie, C.; Li, Z.; Yuan, J.; Wang, X.; Li, Q.; Cheung, S.-C.; Zhang, S.; et al. 2025c. Infi-MMR: Curriculum-based Unlocking Multimodal Reasoning via Phased Reinforcement Learning in Multimodal Small Language Models. *arXiv preprint arXiv:2505.23091*.

Liu, Z.; Sun, Z.; Zang, Y.; Dong, X.; Cao, Y.; Duan, H.; Lin, D.; and Wang, J. 2025d. Visual-RFT: Visual Reinforcement Fine-Tuning. *arXiv:2503.01785*.

Lu, Y.; Yang, J.; Shen, Y.; and Awadallah, A. 2024. OmniParser for Pure Vision Based GUI Agent. *arXiv:2408.00203*.

Lu, Z.; Chai, Y.; Guo, Y.; Yin, X.; Liu, L.; Wang, H.; Xiong, G.; and Li, H. 2025. UI-R1: Enhancing Action Prediction of GUI Agents by Reinforcement Learning. *arXiv preprint arXiv:2503.21620*.

Luo, R.; Wang, L.; He, W.; Chen, L.; Li, J.; and Xia, X. 2025. Gui-r1: A generalist rl-style vision-language action model for gui agents. *arXiv preprint arXiv:2504.10458*.

Nayak, S.; Jian, X.; Lin, K. Q.; Rodriguez, J. A.; Kalsi, M.; Awal, R.; Chapados, N.; Özsü, M. T.; Agrawal, A.; Vazquez, D.; et al. 2025. UI-Vision: A Desktop-centric GUI Benchmark for Visual Perception and Interaction. *arXiv:2503.15661*.

Nguyen, D.; Chen, J.; Wang, Y.; Wu, G.; Park, N.; Hu, Z.; Lyu, H.; Wu, J.; Aponte, R.; Xia, Y.; et al. 2024. GUI Agents: A Survey. *arXiv:2412.13501*.

Pan, J.; Liu, C.; Wu, J.; Liu, F.; Zhu, J.; Li, H. B.; Chen, C.; Ouyang, C.; and Rueckert, D. 2025. MedVLM-R1: Incentivizing Medical Reasoning Capability of Vision-Language Models (VLMs) via Reinforcement Learning. *arXiv:2502.19634*.

Park, J.; Tang, P.; Das, S.; Appalaraju, S.; Singh, K. Y.; Manmatha, R.; and Ghadar, S. 2025. R-VLM: Region-Aware Vision Language Model for Precise GUI Grounding. *arXiv:2507.05673*.

Qin, Y.; Ye, Y.; Fang, J.; Wang, H.; Liang, S.; Tian, S.; Zhang, J.; Li, J.; Li, Y.; Huang, S.; et al. 2025. UI-TARS: Pioneering Automated GUI Interaction with Native Agents. *arXiv preprint arXiv:2501.12326*.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Tang, F.; Gu, Z.; Lu, Z.; Liu, X.; Shen, S.; Meng, C.; Wang, W.; Zhang, W.; Shen, Y.; Lu, W.; et al. 2025. GUI-G<sup>2</sup>: Gaussian Reward Modeling for GUI Grounding. *arXiv:2507.15846*.

Wang, X.; Wu, Z.; Xie, J.; Ding, Z.; Yang, B.; Li, Z.; Liu, Z.; Li, Q.; Dong, X.; Chen, Z.; et al. 2025. MMBench-GUI: Hierarchical Multi-Platform Evaluation Framework for GUI Agents. *arXiv preprint arXiv:2507.19478*.

Wang, Y.; Zhang, H.; Tian, J.; and Tang, Y. 2024. Ponder & Press: Advancing Visual GUI Agent towards General Computer Control. *arXiv:2412.01268*.

Wei, J.; Liu, J.; Liu, L.; Hu, M.; Ning, J.; Li, M.; Yin, W.; He, J.; Liang, X.; Feng, C.; et al. 2025. Learning, Reasoning, Refinement: A Framework for Kahneman’s Dual-System Intelligence in GUI Agents. *arXiv:2506.17913*.

Wu, H.; Chen, H.; Cai, Y.; Liu, C.; Ye, Q.; Yang, M.-H.; and Wang, Y. 2025a. DiMo-GUI: Advancing Test-time Scaling in GUI Grounding via Modality-Aware Visual Reasoning. *arXiv:2507.00008*.

Wu, P.; Ma, S.; Wang, B.; Yu, J.; Lu, L.; and Liu, Z. 2025b. GUI-Reflection: Empowering Multimodal GUI Models with Self-Reflection Behavior. *arXiv:2506.08012*.

Wu, Q.; Cheng, K.; Yang, R.; Zhang, C.; Yang, J.; Jiang, H.; Mu, J.; Peng, B.; Qiao, B.; Tan, R.; et al. 2025c. GUI-Actor: Coordinate-Free Visual Grounding for GUI Agents. *arXiv:2506.03143*.

Wu, Z.; Wu, Z.; Xu, F.; Wang, Y.; Sun, Q.; Jia, C.; Cheng, K.; Ding, Z.; Chen, L.; Liang, P. P.; et al. 2024. OS-ATLAS: A Foundation Action Model for Generalist GUI Agents. *arXiv:2410.23218*.

Xie, C.; Cai, S.; Wang, W.; Li, P.; Sang, Z.; Yang, K.; Zhang, Y.; Li, Z.; Zhu, G.; Liu, Z.; et al. 2025. Infir: Crafting effective small language models and multimodal small language models in reasoning. *arXiv preprint arXiv:2502.11573*.

Xu, Y.; Wang, Z.; Wang, J.; Lu, D.; Xie, T.; Saha, A.; Sahoo, D.; Yu, T.; and Xiong, C. 2025. Aguis: Unified Pure Vision Agents for Autonomous GUI Interaction. *arXiv:2412.04454*.

Yang, J.; Tan, R.; Wu, Q.; Zheng, R.; Peng, B.; Liang, Y.; Gu, Y.; Cai, M.; Ye, S.; Jang, J.; et al. 2025a. Magma: A Foundation Model for Multimodal AI Agents. *arXiv:2502.13130*.

Yang, Y.; Li, D.; Dai, Y.; Yang, Y.; Luo, Z.; Zhao, Z.; Hu, Z.; Huang, J.; Saha, A.; Chen, Z.; et al. 2025b. GTA1: GUI Test-time Scaling Agent. *arXiv preprint arXiv:2507.05791*.

Yang, Y.; Wang, Y.; Li, D.; Luo, Z.; Chen, B.; Huang, C.; and Li, J. 2025c. Aria-UI: Visual Grounding for GUI Instructions. *arXiv:2412.16256*.

Yang, Z.; Li, L.; Lin, K.; Wang, J.; Lin, C.-C.; Liu, Z.; and Wang, L. 2023. The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision). *arXiv:2309.17421*.

Yu, H.; Qu, Z.; Yu, Q.; Chen, J.; Jiang, Z.; Chen, Z.; Zhang, S.; Xu, J.; Wu, F.; Lv, C.; et al. 2024. GaussianTalker: Speaker-specific Talking Head Synthesis via 3D Gaussian Splatting. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, 3548–3557.

Yuan, X.; Zhang, J.; Li, K.; Cai, Z.; Yao, L.; Chen, J.; Wang, E.; Hou, Q.; Chen, J.; Jiang, P.-T.; et al. 2025. Enhancing Visual Grounding for GUI Agents via Self-Evolutionary Reinforcement Learning. *arXiv:2505.12370*.

Zhang, Z.; Lu, Y.; Fu, Y.; Huo, Y.; Yang, S.; Wu, Y.; Si, H.; Cong, X.; Chen, H.; Lin, Y.; et al. 2025. AgentCPM-GUI: Building Mobile-Use Agents with Reinforcement Fine-Tuning. *arXiv:2506.01391*.

Zhou, G.; Qiu, P.; Chen, C.; Wang, J.; Yang, Z.; Xu, J.; and Qiu, M. 2025a. Reinforced MLLM: A Survey on RL-Based Reasoning in Multimodal Large Language Models. *arXiv:2504.21277*.

Zhou, Y.; Dai, S.; Wang, S.; Zhou, K.; Jia, Q.; and Xu, J. 2025b. GUI-G1: Understanding r1-zero-like training for visual grounding in gui agents. *arXiv preprint arXiv:2505.15810*.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. *arXiv:2504.10479*.