

Qwen-UI-Agent Technical Report: Toward Next-Generation Real-World Centric Foundation GUI Agents

MAI-UI Team*

Alibaba Token Hub , Alibaba Group

 tongyi-mai.github.io/Qwen-UI-Agent

Abstract

GUI agents have the potential to become a general purpose executor over existing digital devices. To advance them toward real-world use, we envision agents that operate reliably on real devices, execute workflows across platforms, combine GUI interaction with CLI execution, complete long-horizon tasks, proactively initiate useful services, and autonomously improve their capabilities with minimal human effort. Guided by this vision, we present **Qwen-UI-Agent**, a real-world centric foundation GUI agent spanning mobile, computer-use, web, and DeepSearch environments. Qwen-UI-Agent combines diverse sandbox environments with a large-scale real-device mobile runtime. Its unified action space interleaves GUI operations with CLI execution and generates batched actions in a single model turn. An AutoResearch-style data flywheel uses agents to construct tasks and environments, diagnose failures, and plan subsequent iterations. Online RL support training on trajectories exceeding 100 turns, with over 10,000 concurrent environments accelerating rollout. A lightweight harness layer supports proactive service initiation and stateful workflows across mobile and computer. Across broad evaluations, Qwen-UI-Agent achieves competitive or superior performance against frontier models, including Opus 4.8, Gemini 3.1 Pro, and GPT-5.6 Sol. On mobile use, it achieves **82.1%** on MobileWorld, **92.2%** on MobileWorld-Real, and **97.5%** on AndroidDaily. On computer-use tasks, Qwen-UI-Agent achieves **79.5%** on OSWorld-Verified and a **40.0%** partial-progress score on OSWorld-v2. Across browser use, DeepSearch, and GUI grounding, Qwen-UI-Agent achieves **73.6%** on WebArena, **75.0%** on BrowseComp-ZH, and **81.5%** on ScreenSpot-Pro, respectively.

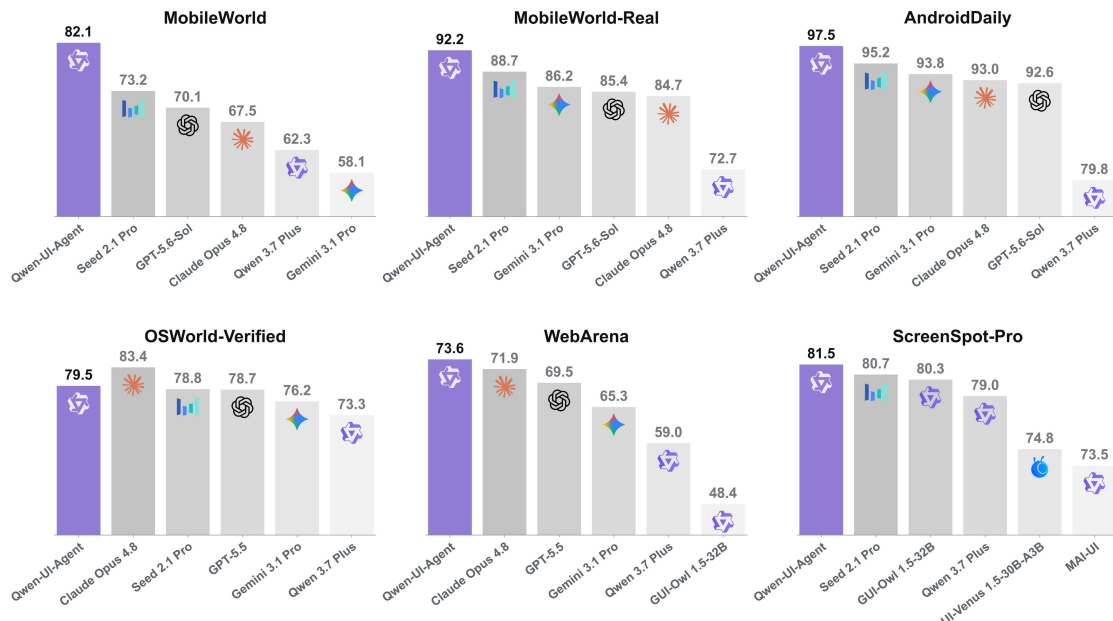


Figure 1: Qwen-UI-Agent demonstrates leading or competitive performance across diverse GUI settings.

*Qwen-UI-Agent is a continuation of our previous work, MAI-UI (Zhou et al., 2025a). See [Contributions](#) section for a full author list.

Contents

1	Introduction	3
2	Qwen-UI-Agent	5
2.1	System Overview	6
2.1.1	Task Formulation	6
2.1.2	Action Space	6
2.1.3	Capability Landscape	7
2.2	Environment Infrastructure	7
2.2.1	Scaling Sandbox Environments	8
2.2.2	Real-Device Mobile Environment: From Simulated Sandboxes to Real-World Learning	9
2.2.3	Hybrid GUI+CLI Action Space	10
2.2.4	Unified Interface for Heterogeneous Environments	11
2.3	Agent-Driven Data Flywheel	11
2.4	Training	13
2.4.1	Supervised Fine-Tuning (SFT)	13
2.4.2	Action RL: Correcting Recurring Action Errors	14
2.4.3	Online RL: Learning Long-Horizon Decision Making	15
2.5	Harness Layer: Proactive Service and Cross-Platform Task Execution	16
2.5.1	Proactive Service Based on Mobile Notifications	17
2.5.2	Cross-Platform Task Execution	19
3	Experiments	20
3.1	Experimental Setup	20
3.2	MobileWorld-Real: Real-world Mobile Use Evaluation Benchmark	21
3.3	Main Results	23
3.3.1	Mobile-Use Evaluation.	23
3.3.2	Computer-Use Evaluation	25
3.3.3	Browser-Use and DeepSearch Evaluation	26
3.3.4	GUI Grounding	29
3.3.5	General and Agentic Capabilities	30
3.4	Harness-Enabled Workflows	31
4	Behavioral Analysis of GUI Agent Execution	32
4.1	How Does a Model’s Limited Real-Device Experience Affect Its Execution Behavior?	32
4.2	A Closer Look at GUI-CLI Coordination and Batched Execution in Qwen-UI-Agent	34
4.3	Step-Level Behavioral Corrections after Action RL	39
4.4	Trajectory-Level Behavioral Shifts after Online RL	41
5	Related Work	42
6	Conclusion	43
7	Limitations and Future Directions	44
8	Contributions	45
A	Appendix	52
A.1	Validating AutoJudge for Real-Device Evaluation	52
A.2	AutoJudge Decisions on Real-Device GUI Agent Trajectories	52
A.3	Additional Qualitative Examples	53

1 Introduction

Graphical user interfaces (GUIs) are the primary interface through which humans access digital services and applications. A capable GUI agent (Zhou et al., 2025a; Wang et al., 2025a; Qin et al., 2025; Xu et al., 2026) therefore has the potential to become a general purpose executor over existing application ecosystems: it can understand user intent, perceive application states, operate interfaces, and complete tasks without requiring every service to expose a dedicated API. Recent progress in multimodal foundation models has made this direction increasingly practical, leading to rapid improvements in mobile and desktop GUI agents (Cao et al., 2026; Gao et al., 2026; Huang et al., 2026; Shen et al., 2026).

Despite this progress, GUI agents remain largely optimized for simulated benchmarks, leaving a substantial gap between benchmark performance and real-world utility. To bridge this gap, we envision a next generation of GUI agents that operate reliably on real devices, complete long-horizon tasks, carry workflows across platforms, proactively initiate useful services, combine GUI interaction with CLI execution, and continuously evolve their capabilities with less human effort. To realize this vision, we identify six key transitions: **(1) From simulated environments to real-device execution.** Many GUI agents, especially mobile agents, are optimized for simulated benchmarks, while real mobile devices involve much more complex application functions, varied UI layouts, interruptions, permissions, and changing environments. A model that performs well in simulation often fails on real devices. **(2) From isolated domains to cross-domain, cross-platform workflows.** Existing agents are often built for mobile, web, or desktop separately, while real-world tasks may span several domains and continue across phones and computers. **(3) From GUI-only actions to hybrid GUI+CLI and batched actions.** GUI interaction provides universal access to visual interfaces, while the growing coding capabilities of foundation models make CLI a natural complement for structured tasks such as data processing and file manipulation. By combining these action types and supporting batched actions, agents can better match the execution method to the nature of tasks, improving execution efficiency and expanding task coverage. **(4) From short-horizon tasks to reliable long-horizon task completion.** The practical value of a GUI agent grows with the complexity of the user objectives it can complete and the number of execution steps it can perform on the user’s behalf. However, completing long-horizon tasks requires sustained planning, state tracking, intermediate verification, and recovery from execution errors, making long-horizon reliability a challenging capability. **(5) From human-intensive training pipelines to AutoResearch-style one.** Current GUI agent development still relies heavily on manually constructing tasks, collecting data, training models, running evaluations, and analyzing failures. As agent capabilities improve, this process should become increasingly automated. **(6) From reactive execution to proactive service initiation.** Most agents wait for explicit user instructions, even though many useful tasks can arise directly from digital signals on phones or other devices. An agent that recognizes such actionable moments and proposes appropriate next steps can provide a fundamentally different user experience.

In this report, we introduce Qwen-UI-Agent, a real-world-centric foundation GUI agent developed around these five transitions, representing our step toward next-generation GUI agents for real-world utility. Specifically, Qwen-UI-Agent delivers the following core features:

- **Real-device Mobile GUI Foundation.** We find that the simulation-to-real gap is particularly severe in mobile environments. To address this gap, we build a real-device mobile environment comprising over 100 physical devices and supporting more than 150 applications. We use it for realistic task design, trajectory collection, online reinforcement learning, and evaluation on real devices. To improve practical usability, we further support user takeover for high-risk actions, expand coverage to both high-frequency workflows and less commonly used app functions. Together, these efforts enable Qwen-UI-Agent to substantially outperform existing models on real phones.
- **Multi-domain Capabilities and Cross-Domain Workflows.** Qwen-UI-Agent supports task execution across mobile, web, computer-use, and DeepSearch environments. Rather than treating these capabilities as isolated modules, our harness layer composes them within a unified workflow, preserving

context and task state as execution moves seamlessly across devices and environments. For example, GUI workflows can invoke DeepSearch to efficiently retrieve and verify external information, avoiding time-consuming GUI-based web navigation before seamlessly resuming interaction. This turns multi-domain coverage into integrated task execution across heterogeneous digital environments.

- **Hybrid GUI+CLI and Batched Action Spaces.** We equip Qwen-UI-Agent with a unified action space that supports GUI operations, bash-based CLI commands, and batched actions. The model can select and compose these actions within a single trajectory, using GUI interaction and CLI execution where each is most suitable, while batching compatible actions to reduce unnecessary steps. In computer-use tasks, CLI commands and GUI clicks emerge as the two dominant action types, and more than 30% of action outputs are batched, substantially shortening execution trajectories.
- **Scalable Online RL for Long-horizon Tasks.** Hybrid GUI+CLI and batched actions improve local execution efficiency, but completing long-horizon tasks also requires reliable planning, state tracking, and intermediate decision-making across the full trajectory. We therefore scale verifier-guided online RL to trajectories exceeding 100 interaction steps. We further enable approximately 10,000 simulated environments to run concurrently to accelerate rollout generation while a model-adaptive curriculum prioritizes tasks with intermediate success rates and replaces mastered tasks with harder ones.
- **AutoResearch-style Capability Acquisition.** We build an agent-driven iteration loop in which agents generate and assess candidate tasks for data flywheel, process rollout trajectories, analyze failures, and propose plans for the next iteration. Rather than relying on humans to manually drive every stage, the system automates much of the capability development process, with humans mainly providing supervision and making targeted revisions. This approach reduces human effort and accelerates the acquisition and refinement of new GUI agent capabilities.
- **Proactive Mobile Service.** Consider a flight cancellation notification arriving on a user’s phone. Instead of waiting for the user to notice it and issue an instruction, the agent can identify the affected itinerary, gather alternative travel options, check against user’s schedule, and present a proposed plan for confirmation. To support such experiences, we build a proactive service harness based on mobile notifications. Beyond initiating tasks, the harness forms a coherent understanding of real-world situations, generates personalized tasks, executes approved actions, and continuously evolves from user responses and long-term interaction history.

Empirically, Qwen-UI-Agent achieves state-of-the-art performance across a broad range of evaluation settings, covering real-device and benchmark-based mobile use, computer use, browser use, DeepSearch, and GUI grounding. Across these evaluations, Qwen-UI-Agent generally outperforms leading foundation models, including Opus 4.8 ([Anthropic, 2026](#)), GPT-5.6 Sol ([OpenAI, 2026b](#)), Gemini 3.5 Flash ([Google, 2026b](#)), and Seed2.1 Pro ([ByteDance Seed, 2026b](#)), as well as specialized GUI agents. We highlight main results below:

- **Mobile Use.** On MobileWorld-Real, our real-device benchmark comprising more than 400 tasks across over 100 apps, Qwen-UI-Agent achieves a **92.2%** success rate. It outperforms frontier closed-source models Gemini 3.1 Pro, Claude Opus 4.8, GPT-5.6 Sol, and Seed 2.1 Pro by 6.0, 7.5, 6.8, and 3.5 percentage points, respectively. Qwen-UI-Agent further achieves a near-perfect success rate of **97.5%** on AndroidDaily, another real-device benchmark. On MobileWorld, Qwen-UI-Agent scores **82.1%**, surpassing Opus 4.8, GPT-5.6 Sol, and Seed 2.1 Pro by 14.6, 12.0, and 8.9 percentage points.
- **Computer Use.** On OSWorld-Verified, Qwen-UI-Agent achieves **79.5%**, ranking second overall and outperforming GPT-5.5, Gemini 3.1 Pro, and Seed 2.1 Pro. On the more challenging OSWorld-v2, Qwen-UI-Agent obtains the second-best binary success rate (**13.9%**), the third-best partial-progress score (**40.0%**). Its partial-progress score exceeds those of MiniMax M3 and Qwen 3.7 Plus by **17.7** and **18.5** percentage points, respectively, while requiring **58.4%** and **21.7%** fewer steps per task.
- **Browser Use and DeepSearch.** On WebArena, Qwen-UI-Agent achieves **73.6%**, outperforming Claude Opus 4.8, GPT-5.5 and Gemini 3.1 Pro by 1.7, 4.1 and 8.3 points. On DeepSearch, it scores **64.1%** on

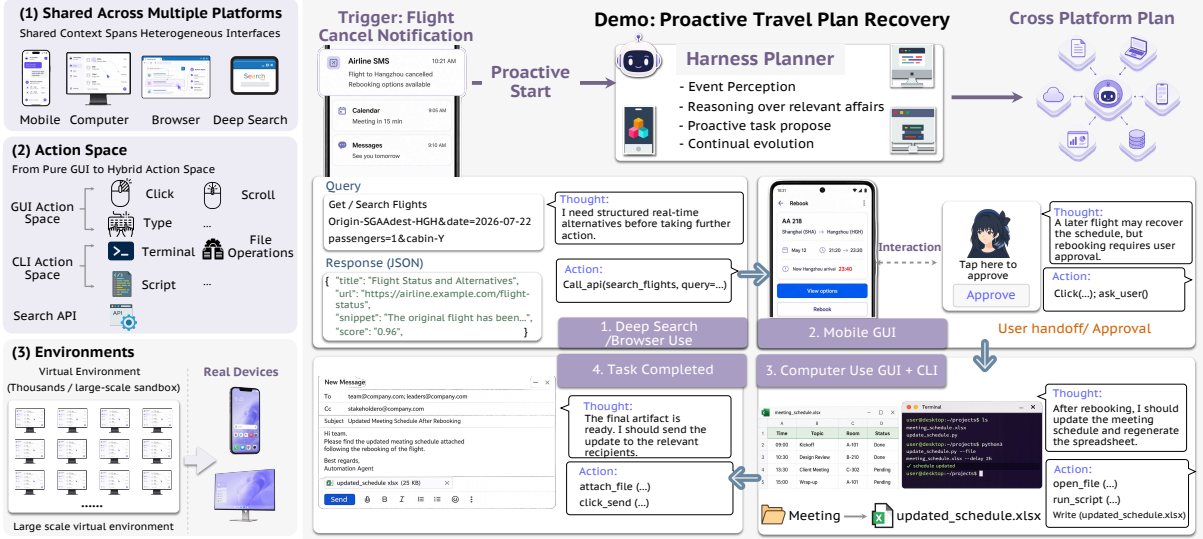


Figure 2: An illustrative trajectory of Qwen-UI-Agent for proactive cross-platform task execution. The left panel summarizes the system capabilities underlying the trajectory, including cross-platform execution, a hybrid action space spanning GUI, CLI, and API operations, and environments ranging from large-scale sandboxes to real devices. The right panel illustrates a travel-recovery scenario triggered by a flight-cancellation notification. After identifying the affected tasks and commitments, the agent searches for alternative flights through an API, requests user approval before rebooking via a mobile GUI, updates the affected meeting schedule through desktop GUI and CLI actions, and sends the revised artifact to the relevant recipients.

BrowseComp and 75.0% on BrowseComp-ZH, surpassing Qwen3.5-397B-A17B and UI-TARS-2.

- **GUI Grounding.** Qwen-UI-Agent reaches 81.5% on ScreenSpot-Pro under the zoom-in setting, outperforming Seed 2.1 Pro, GUI-Owl-1.5, and UI-Venus-1.5. It further sets the best results among baseline models on UI-Vision (70.0%), OSWorld-G-Refined (78.5%), MMBench-GUI L2 (92.6%), and ScreenSpot-V2 (97.5%).
- **General and Agentic Capabilities.** To preserve the model’s broad utility in real-world settings, we retain Qwen-UI-Agent’s general reasoning and agentic capabilities. We evaluate Qwen-UI-Agent on a suite of 13 benchmarks, including MMMU-Pro, MMLU-Pro, Terminal-Bench 2.0, and Claw-Eval. The results show that Qwen-UI-Agent outperforms the Qwen base model on agentic tasks while remaining comparable on general reasoning tasks, and significantly outperforms GUI-specialized models across both categories.

2 Qwen-UI-Agent

This section presents the methodology of Qwen-UI-Agent as an integrated system for building advanced foundation GUI agents. The system comprises four components: (1) an environment infrastructure spanning both sandboxed and real-world environments that supports trajectory collection and training across mobile-use, computer-use, browser-use, and deep-search tasks under a unified action space covering GUI, CLI, and batched actions; (2) an automated data flywheel that reduces human intervention and closes the loop from data synthesis to training, evaluation, failure analysis, and targeted data generation for the next iteration; (3) a unified training framework that combines supervised fine-tuning with reinforcement learning; and (4) a harness layer that connects the trained agent to user context and enables proactive services and cross-platform task execution. Figure 2 illustrates a demo trajectory.

2.1 System Overview

Qwen-UI-Agent is designed as a real-world-centric foundation GUI agent that operates across diverse digital environments, including mobile, desktop, web, and information-seeking systems. This section first defines the task scope of Qwen-UI-Agent, then introduces its unified observation and action spaces, and finally summarizes the capability landscape supported by this formulation.

2.1.1 Task Formulation

Qwen-UI-Agent is designed for interactive digital task execution. We define a task as:

$$\tau = (I, \mathcal{E}_\tau), \quad (1)$$

where I is the user instruction and $\mathcal{E}_\tau$ is the set of digital environments available to the task. Depending on the task, $\mathcal{E}_\tau$ may include mobile devices, web browsers, computer systems, DeepSearch systems, or a combination of them. This formulation covers both tasks completed within a single environment and workflows that span multiple platforms (Xu et al., 2026; Lian et al., 2026).

At decision step t , the agent receives a multichannel observation o_t that may combine different forms of environment feedback:

$$o_t = (o_t^{\text{GUI}}, o_t^{\text{CLI}}, o_t^{\text{API}}). \quad (2)$$

Here, o_t^{GUI} represents the current screenshot, o_t^{CLI} contains command execution results, and o_t^{API} contains structured responses from external services, such as search API. A Component is set to empty when the corresponding channel is unavailable or unnecessary for the current task. This multichannel observation allows the agent to jointly reason over visual interface states and programmatic execution feedback (Hu et al., 2026; Li et al., 2026a).

The agent predicts an intermediate reasoning output r_t and an executable action output $\mathbf{a}_t$:

$$(r_t, \mathbf{a}_t) = \pi_\theta(I, o_t, h_t), \quad (3)$$

where $h_t = (o_1, r_1, \mathbf{a}_1, \dots, o_{t-1}, r_{t-1}, \mathbf{a}_{t-1})$ denotes the preceding interaction history.

Notably, one model decision step may produce either a single action or an ordered sequence of actions:

$$\mathbf{a}_t = (a_t^{(1)}, \dots, a_t^{(K_t)}), \quad a_t^{(k)} \in \mathcal{A}_t, \quad (4)$$

where $\mathcal{A}_t$ is the set of actions available in the current environment. When $K_t = 1$, the model performs single-action execution. When $K_t > 1$, it produces a batched action sequence. Actions within a batch are executed consecutively in this decision step, reducing unnecessary inference and observation steps when multiple operations can be completed without additional environment feedback.

2.1.2 Action Space

We design the action space of Qwen-UI-Agent around three requirements: supporting GUI interaction across different platforms, extending execution beyond GUI manipulation, and preserving user control over consequential operations. The GUI action space is defined as the union of actions necessary for mobile, web, and desktop environments, with each environment exposing the subset supported by its native interaction mechanisms. To complement GUI interaction, we introduce `cli_command`, which enables direct bash command execution, and `api_call`, which invokes external services with structured arguments. We further include `ask_user`, allowing the agent to request missing information or obtain explicit confirmation before handling sensitive data, payments, or other consequential operations. Together, these actions provide a unified interface for cross-platform GUI interaction, hybrid GUI+CLI execution, and user-controlled operation in real-world tasks. The full action space is shown in Table 1.

Table 1: Action Space in Qwen-UI-Agent.

Action	Definition
GUI Actions	
click	Clicks at coordinates (x, y) .
double_click	Double-clicks at coordinates (x, y) .
long_press	Long-presses at coordinates (x, y) .
type	Types the specified text content.
open	Opens the specified app.
drag	Drags from start coordinates (x_1, y_1) to end coordinates (x_2, y_2) .
system_button	Presses a system button, selected from back, home, menu, and enter.
wait	Waits for a specified duration in seconds.
CLI Actions	
cli_command	Executes a bash command in the active CLI environment.
API Actions	
api_call	Invokes an API with specified arguments.
Interaction and Control Actions	
ask_user	Interact with the user to complete the task.
terminate	Ends the task with a final answer and a status of success or failed.

2.1.3 Capability Landscape

The unified formulation supports four complementary capabilities of Qwen-UI-Agent centered on real-world task execution. (1) **Real-device mobile execution.** By building a real-device mobile environment and using it throughout model capability development and evaluation, Qwen-UI-Agent substantially eliminates the simulation-to-reality gap and achieves leading real-world mobile execution performance over both specialized GUI agents and frontier closed-source models. (2) **Multi-domain task execution.** Qwen-UI-Agent operates across mobile, desktop, web, and DeepSearch environments, moving beyond domain-specific policies toward a foundation GUI agent for heterogeneous digital tasks. (3) **Hybrid and efficient interaction.** Qwen-UI-Agent’s unified action space enables the model to select and interleave GUI operations with bash-based CLI commands within the same trajectory, while batched actions largely enhance the execution efficiency. (4) **Harness-enabled services.** A lightweight harness layer extends the core agent beyond isolated task execution by enabling proactive service initiation and allowing a single workflow to span mobile and computer systems.

2.2 Environment Infrastructure

The environment defines the capability boundary of an agent: it determines what the agent can perceive, what actions it can take, and what experience it can acquire through interaction. Therefore, moving GUI agents from benchmark-centric optimization toward real-world utility requires not only stronger models, but also broader and more realistic environments in which agents can learn, act, and adapt. As illustrated in Figure 3, our environment infrastructure is organized around four core components. First, **simulated and sandbox environments** provide scalable, controllable, and reproducible settings for data generation, agent learning, and evaluation. Second, extending these simulated environments to **real devices** enables agents to learn and operate under realistic deployment conditions, including dynamic device states, system constraints, and user interaction. Third, a **hybrid GUI and CLI action space** supports more efficient and adaptive execution by allowing the agent to select the most appropriate interaction modality for each task. Finally, a **unified cross-platform environment** provides a unified pipeline across mobile-use, computer-use, browser-use, and DeepSearch, while preserving context and task state across platforms. Together, these components make the environment a core part of the agent system, directly shaping the range of tasks the agent can learn and accomplish.

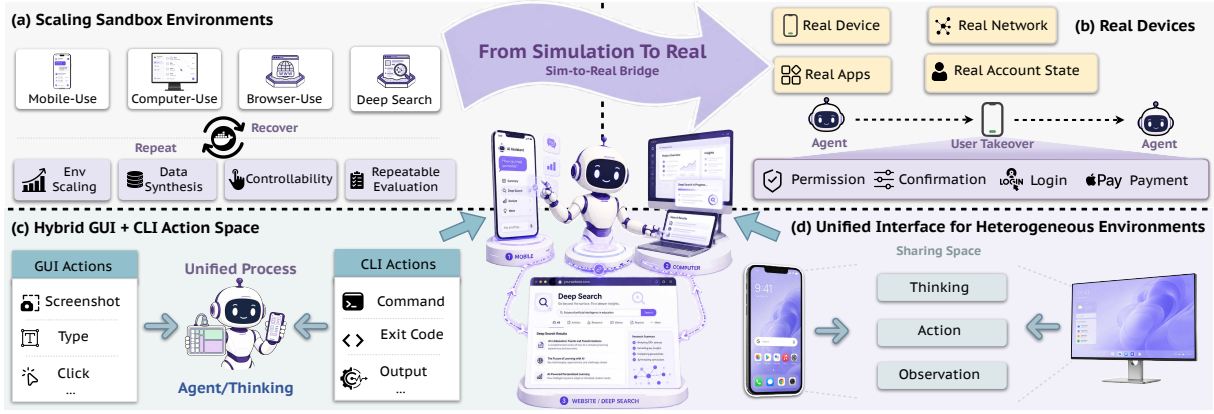


Figure 3: The environment infrastructure of Qwen-UI-Agent. (a) Scalable sandbox environments spanning mobile-use, computer-use, browser-use, and DeepSearch provide controllability and repeatable evaluation for data synthesis and training. (b) A sim-to-real bridge extends agents to real devices with real applications, networks, and account states, supporting user takeover for login, payment, permission, and confirmation. (c) A hybrid GUI+CLI action space interleaves graphical operations with direct command execution. (d) A unified interface standardizes the thinking–action–observation loop across heterogeneous environments.

2.2.1 Scaling Sandbox Environments

Sandbox environments provide controlled and resettable task states for training, evaluation, and ablation. Our suite spans mobile-use, computer-use, browser-use, and DeepSearch tasks. Each domain retains its native applications, environment-specific context, and state-based verification mechanisms. A key strength of our infrastructure is its scalability: it supports up to **10,000** isolated sandbox environments running in parallel, enabling large-scale rollout and agent training.

- **Mobile Use.** We build our mobile sandbox on the environment provided by MobileWorld, which offers a resettable Android system with 20 applications and supports GUI task execution, agent-user interaction, and MCP tool use (Kong et al., 2026). To address the scaling bottleneck of KVM-based emulators, we rebuilt the MobileWorld environment on redroid (Remote Android Project, 2024), which runs Android as a container on the host kernel without QEMU or nested KVM. This design enables complete device-and-backend environments to be replicated across ordinary container hosts. Its containerized device instances and self-hosted application backends provide controllable initial states and reproducible state transitions. Building on this infrastructure, we construct diverse task-relevant environment contexts, synthesize tasks, and generate state-based verifiers. The resulting sandbox provides a stable and scalable environment for trajectory collection and agent training.
- **Computer Use.** We build our computer-use sandbox on the Ubuntu virtual-machine environment provided by OSWorld, which supports real desktop applications, operating-system operations, and cross-application workflows (Xie et al., 2024). While the standard OSWorld interaction interface is GUI-based, we extend it with direct bash execution, allowing the agent to interleave GUI operations and CLI commands within the same trajectory. For our newly constructed tasks, we further create task-specific initial environment states and state-based verifiers for some of these tasks. Together, these extensions provide a controllable and scalable environment for collecting trajectories and training agents under both GUI-only and hybrid GUI+CLI interaction.
- **Browser Use.** Our web environment is a self-contained browser runtime implemented with FastAPI, Playwright (Microsoft, 2020), and Chromium. Each episode runs in a fresh Playwright BrowserContext with isolated cookies, cache, and local storage, and begins from a task-specific page and state. The agent receives rendered screenshots and interacts through browser-native GUI actions. The JavaScript-based verifiers inspect the DOM and persistent application state, allowing alternative action sequences to

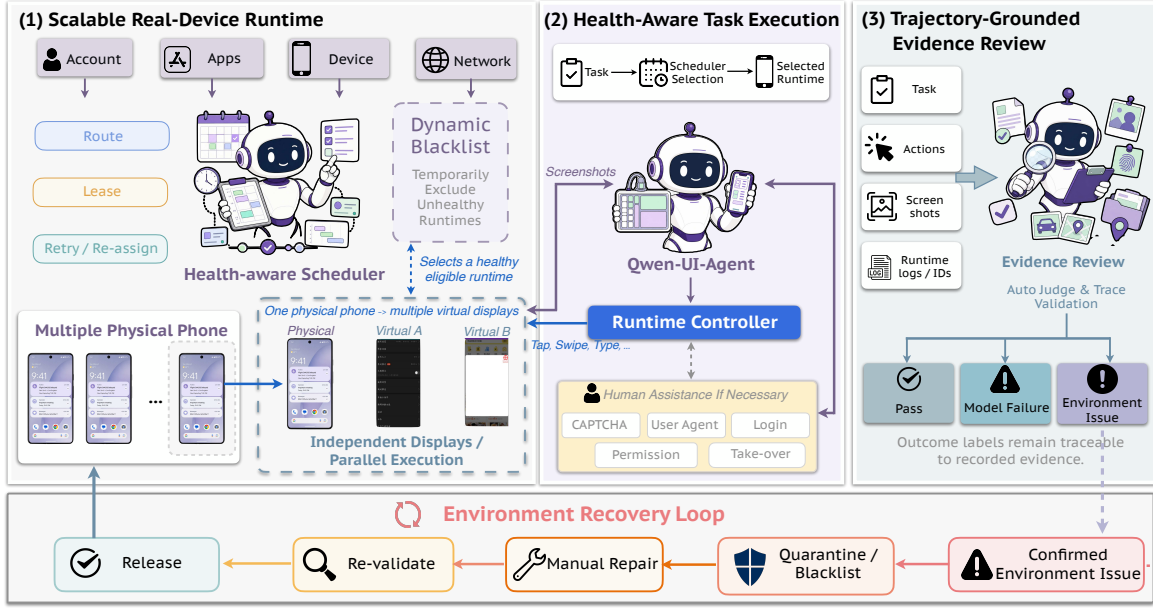


Figure 4: Real-device mobile runtime with closed-loop environment governance. The health-aware scheduler routes each task to an eligible phone, App/account, and display; unhealthy targets remain blacklisted until repair and revalidation. Virtual displays allow one phone to run multiple Apps concurrently. Evidence-based review separates task success, model failure, and environment failure from the complete trajectory, and confirmed environment issues are fed back to the scheduler.

receive credit when they reach the required state.

- **DeepSearch.** The DeepSearch environment complements interactive browser control with structured information-access tools. Serper (Serper, 2023) provides ranked search results, while Jina Reader (Jina AI, 2024) converts selected web pages and documents into model-consumable text. The agent can iteratively reformulate queries, select sources, retrieve full content, and synthesize evidence across pages. This environment therefore targets open-ended, multi-source information seeking in which the principal observations are search results and retrieved documents rather than rendered GUI states.

2.2.2 Real-Device Mobile Environment: From Simulated Sandboxes to Real-World Learning

Why Real Devices Matter? Sandbox environments provide stable and resettable task states, making them essential for large-scale rollout and training. However, they cannot fully capture the evolving application states, permission-related constraints, and or execution-time disruptions such as unexpected pop-ups and network instability. On real devices, application interfaces and content change over time, permissions and CAPTCHA interrupt execution, and network or device failures can occur at any step. The Chinese mobile ecosystem further amplifies these challenges: super Apps, dense interfaces, and frequent pop-ups make it impossible to reproduce such execution paths in a sandbox. To make Qwen-UI-Agent work reliably under these conditions, we build a real-device mobile runtime and use it for model training and evaluation. We further construct MobileWorld-Real, which is introduced in Section 3.2, to systematically evaluate agent performance on real devices.

Real-device Mobile Runtime. We build a real-device mobile runtime comprising over 100 physical devices and over 150 applications. At this scale, assigning each task a reliable and usable execution device becomes a central systems challenge. As shown in Figure 4, the health-aware scheduler continuously tracks the health and availability of each device, application, account, network connection, and display.

When a task enters the queue, it selects an eligible execution target, leases the required resources to the task, and reroutes execution to another target when failures occur. This health-aware orchestration keeps the environment productive despite environment-side failures caused by the inherent instability of real phones, applications, and network connections.

To further scale real-device environment, we leverage the virtual-screen mechanism ([Genymobile, 2018](#)) that allows a single phone to host multiple application sessions concurrently on separate displays. The runtime controller binds each virtual display to a corresponding agent session, ensuring that observations are delivered to the correct agent and that its actions are executed on the intended display. This design increases the number of active execution environments without requiring a proportional increase in physical devices. Additionally, real-world tasks often involve either sensitive operations that require explicit user approval or missing information that must be provided by the user. We therefore introduce a dedicated User Agent that can provide missing information, obtain explicit confirmation for data-sensitive or payment-related operations, and hand control to the user when an action requires direct human intervention such as CAPTCHA. After the required input or operation is completed, the task resumes from the updated environment state.

A real-device runtime also needs to distinguish model failures from failures of the execution environment itself. Application restrictions, unavailable services, unstable devices, and network errors may interrupt a trajectory even when the model’s decisions are valid. The scheduler maintains a dynamic blacklist over unhealthy devices, removes them from routing, and restores them only after manual inspection. We employ a VLM-based judge to examine the complete trajectory and to distinguish among task success, model failure, and environment failure. Confirmed environment failures update the Scheduler and maintenance queue, while validated trajectories are retained for training.

2.2.3 Hybrid GUI+CLI Action Space

Real-device environments expand the range of states an agent can encounter, but real-world utility also depends on how effectively the agent can act on them. GUI interaction provides universal access to user-facing applications and visual context, but pure GUI execution may turn structured operations into long sequences of grounding, clicking, and typing. CLI execution offers a direct programmatic access to files, code, system settings, and batch operations, but is less suitable when tasks depend on visual understanding, application-specific interfaces, or services without programmable access. GUI and CLI therefore offer complementary forms of environment access: GUI enables broad application coverage and visually grounded interaction, while CLI provides efficient execution over programmatic state.

We therefore equip Qwen-UI-Agent with the unified GUI+CLI action space defined in Section 2.1.2, and our environments expose both forms of access through a single execution interface. In the computer-use environment, every action is dispatched to a lightweight execution service inside the VM: GUI actions are translated into native input events, while `cli_command` runs the command in a non-interactive shell, without visually locating and operating a terminal application. The resulting stdout, stderr, and exit status are returned as structured observations alongside the post-action screenshot, allowing the model to condition its next decision jointly on programmatic and visual state.

Reliable CLI execution requires environment-side handling beyond spawning a shell. Command execution is bounded by a timeout budgeted for slow operations such as dependency installation, while failures like non-zero exits and timeouts are returned as error observations rather than aborting the episode, allowing the model to diagnose and recover within the same trajectory. In batched actions, GUI and CLI sub-actions execute in order with their CLI outputs aggregated into a single observation. Executed commands are also recorded into the shell history, so that verifiers inspecting terminal state credit CLI-based solutions as well.

2.2.4 Unified Interface for Heterogeneous Environments

Mobile-use, computer-use, browser-use, and DeepSearch environments expose different resource models, action spaces, observation formats, and verification procedures. We therefore introduce a unified environment interface that standardizes the environment lifecycle and agent-facing inputs and outputs, while delegating execution, reset, and verification to environment-specific adapters. This allows the same agent and training pipeline to operate across heterogeneous environments without imposing identical runtime semantics.

The interface exposes a common asynchronous lifecycle comprising `acquire`, `reset`, `step`, `evaluate`, `tear_down`, and `release`. Each acquired environment is represented as a leased session bound to a backend and, where necessary, a display identifier, which consistently routes screenshots, actions, and evaluations to the same isolated context. Backend adapters implement resource initialization, observation capture, action execution, and evaluation, while task adapters preserve platform-specific reset and verification logic. All environments return a common transition format containing observations, a termination signal, and reward information.

The policy and environment independently declare their native action spaces. When they differ, supported GUI operations are translated through a canonical intermediate representation, with coordinates normalized according to the target platform. The hybrid GUI, CLI, and API actions defined in Section 2.1.2 remain available through native adapters, and their visual or structured results are returned through the same observation stream. This design provides a shared rollout, data-collection, and evaluation pipeline while retaining the native interaction semantics of each platform.

2.3 Agent-Driven Data Flywheel

Continually improving GUI agents requires an iterative, closed-loop process spanning task design, environment construction, verifier development, data curation, model training, evaluation, and error analysis. When humans drive these stages, iteration becomes slow and difficult to scale as the coverage expands across applications, platforms, and capability dimensions. To address this issue, we therefore introduce an AutoResearch-style, agent-driven data flywheel that shifts human involvement from executing each stage to high-level oversight and targeted intervention.

Overall Procedure. As illustrated in Figure 5, we organize our agent driven data flywheel into two stages: domain capability bootstrapping and an iterative refinement loop. During bootstrapping, we use strong foundation models to analyze the domain knowledge and capabilities required for the various domain covering mobile, desktop and web. Guided by this analysis, the agents generate an initial pool of tasks and corresponding environment contexts, from which candidate trajectories are collected. Following MAI-UI (Zhou et al., 2025a), we first perform multiple rounds of rejection sampling and aggregate the accepted trajectories into a unified SFT corpus. This process yields a strong initial policy with broad coverage of the target domain. Building on this policy, our agent-driven failure analysis identifies the model’s remaining weaknesses and maps them to targeted goals for the next cycle. These goals guide the generation of new tasks, environment configurations, and, where applicable, task-specific verifiers. The accepted trajectories are then incorporated into the training corpus to further improve the model, thereby closing the loop and initiating the next iteration. Across both the initial policy construction and iterative improvement stages, strong foundation models generate tasks and environment configurations, synthesize task-specific verifiers, evaluate trajectories, analyze failures, and plan the optimization and execution of subsequent iterations. Human involvement primarily focuses on designing and implementing the overall system, overseeing its operation, and intervening or revising the workflow when necessary. We next describe four design choices that make this agent-driven data flywheel scalable and effective.

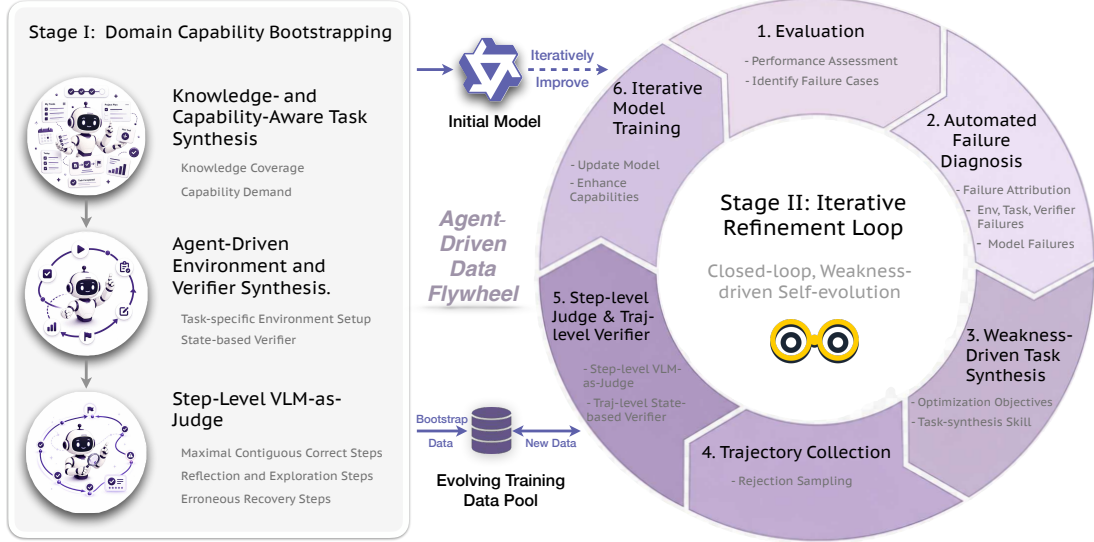


Figure 5: The data flywheel of Qwen-UI-Agent. Domain capability bootstrapping initializes training, iterative refinement loop identifies capability weaknesses and generate targeted tasks, and the resulting data improve the next training iteration.

Knowledge- and Capability-Aware Task Synthesis. Training tasks define the learning signals available to the model and therefore shape the knowledge and behaviors it can acquire. We organize task synthesis along two conceptually distinct dimensions: *knowledge coverage*, which determines what the agent needs to know, and *capability demand*, which determines how the agent must reason and act to complete the task. Knowledge coverage specifies the domain-specific operational knowledge, including application functions, interface conventions, tool usage, and common workflows. Capability demand captures the transferable reasoning and interaction abilities, such as long-horizon state tracking, precise constraint following, numerical and temporal reasoning, and error recovery. For each target domain, we use strong foundation models to construct an initial hierarchical function tree for knowledge coverage and a capability profile for capability demand. We then compile them into a reusable task-synthesis skill containing composition rules, difficulty controls, environment prerequisites, and representative examples. Conditioned on the objective of the current iteration, this skill lets the agent automatically generate task toward the required knowledge and capabilities.

Agent-Driven Environment and Verifier Synthesis. Task synthesis determines the capabilities that a training instance should elicit, but interactive training additionally requires a reproducible environment in which the task is feasible and executable. Given a synthesized task, our agents use *Environment State Synthesis*, detailed in Section 2.4.3, to construct the application state, files, records, accounts, and cross-application context needed for execution. For tasks with reliably checkable outcomes, agents further synthesize state-based verifiers that inspect the resulting application, file, database, or system state, yielding executable task–verifier pairs.

Step-Level Judging for Scalable SFT and Executable Verification for Online RL. Executable verifiers provide high-precision outcome signals, but constructing and validating one is costly. Even with stronger foundation models, we find that synthesizing and validating executable verifiers remains time-consuming, whereas the accuracy of model-based judges improves alongside model development. We therefore extend the step-level trajectory judge used in MAI-UI (Zhou et al., 2025a) and apply a VLM-based judge to mobile, web, and computer-use trajectories, including hybrid GUI+CLI interactions. Given the instruction, trajectory, and visual or structured observations, the judge extracts three forms of useful supervision: maximal contiguous steps that correctly advance the task, the first step that initiates each reflection or exploration phase, and recovery segments that return execution from an erroneous state to a

valid path. Retaining only the first reflection or exploration step captures the useful decision to reconsider without over-keeping a potentially noisy branch. This fine-grained procedure recovers effective training signal from both successful and failed trajectories. Empirically, we find that step-level data filtering yields SFT performance comparable to or better than training on complete trajectories selected by executable verifiers. We therefore use fine-grained VLM judgment to scale process supervision for SFT, while constructing executable verifiers for the high-precision outcome signals required by online RL.

Failure Analysis Driven Iteration. We construct a broad and diagnostic evaluation suite for each domain to track model performance and define the optimization objectives of the next iteration. For each failed task, an analysis agent examines the complete trajectory and distinguishes model failures from environment, task, or verifier failures. Model failures are then mapped to a structured cause, such as missing application knowledge, constraint violations, state-tracking errors, or insufficient verification. The system aggregates these task-level diagnoses to identify the most prevalent and consequential weaknesses of the current policy and converts them into a prioritized set of optimization objectives. Conditioned on these objectives and the task-synthesis skill, agents generate targeted tasks and corresponding environment setups and, where applicable, task-specific verifiers for the next cycle. Non-model issues are routed to environment maintenance or task and verifier revision, where agents diagnose and address the identified problems accordingly.

2.4 Training

2.4.1 Supervised Fine-Tuning (SFT)

Qwen-UI-Agent is designed to support task execution across mobile, computer-use, web, and DeepSearch settings. This scope poses two coupled challenges for SFT: jointly acquiring strong capabilities across all target domains within a single model, and preserving the general reasoning and agentic competence needed for robust real-world task completion.

Domain-Conditioned Expert Training. Across mobile, desktop, and web, each domain places different demands on the agent. Mobile tasks require reliable execution in real-device environments, where interfaces are app-specific, workflows are frequently interrupted by pop-ups, and safety-sensitive actions must remain under user control. Desktop tasks often combine GUI control with Bash-based file and system operations. Web tasks require interaction with dynamic webpages, while DeepSearch tasks require the agent to retrieve, verify, and synthesize evidence through Search APIs.

We train a specialized expert for each domain and then consolidate their capabilities. Each expert is trained mainly on data from its target domain, together with a controlled mixture of data from other domains. This cross-domain mixture helps each expert retain capabilities that transfer across domains and reduces overfitting to its target domain. Practical deployment requires a single model that can operate across all supported settings. We therefore merge the domain-expert checkpoints into a unified model. Model merging combines their complementary strengths within a single checkpoint, preserving the capabilities acquired by the individual experts with little additional training overhead.

Preserving General and Agentic Capabilities with In-Distribution Data. General reasoning and agentic capabilities are essential for real-world use, where agents need to respond to diverse and open-ended user requests. Some requests may even be unrelated to GUI tasks, yet the model should still respond appropriately. Even within GUI tasks, real-world deployment exposes the agent to out-of-distribution (OOD) states and objectives that may require knowledge, reasoning, instruction following, retrieval, coding, or tool use not anticipated during training. We therefore seek to preserve and, where possible, strengthen the general reasoning and agentic capabilities inherited from the base model, providing an important foundation for handling unfamiliar situations and maintaining a consistent user experience.

We find that the most effective strategy is to use in-distribution data that reinforces successful behaviors that the starting model can produce. We construct a broad query pool covering general question

answering, mathematics, coding, visual understanding, search, and agentic tool use, sample responses from the starting model, and retain verified correct examples after format normalization. A small proportion of these examples is mixed with GUI trajectories during SFT. In our experiments, examples that the starting model is capable of solving are substantially more effective for capability preservation than challenging examples on which it fails. The former rehearse established capabilities using targets aligned with the model’s own response distribution, whereas the latter shift training toward capability acquisition and can introduce competing optimization signals.

Sliding-Window Training for Efficient Long-Trajectory SFT. Long-trajectory SFT becomes expensive when every action is packed with its visual and textual context as a separate training sample, causing highly overlapping context to be processed repeatedly across adjacent actions. We therefore divide each trajectory into windows of $n = 5$ consecutive steps and advance each window by $n - 1 = 4$ steps, leaving a one-step overlap between adjacent windows. This overlap ensures that the first newly supervised action in each subsequent window can condition on at least two screenshots. In each subsequent window, the shared boundary step remains in the input context, but its loss is masked because it has already been supervised in the preceding window. Each full subsequent window therefore supervises four new actions. Screenshots before the current window are omitted from the visual input, while the earlier textual trajectory history is retained as context. Joint supervision reduces this repeated context processing and increases the number of supervised action targets per processed context, thereby improving training efficiency.

2.4.2 Action RL: Correcting Recurring Action Errors

A single local error in a long-horizon GUI task can move the interface into an incorrect state and compromise subsequent decisions. SFT learns from successful demonstrations but does not explicitly suppress recurring action errors. We therefore introduce **Action RL**, an action-level reinforcement learning stage that combines targeted error-pattern data with action-aware rewards to improve local decision reliability.

Recurring Action-Error Patterns. Across applications, we identify six recurring patterns:

- *Confusable-Element Grounding.* When a screen contains visually or semantically similar icons, controls, or list items, the model may ground the action to a nearby distractor rather than the correct one.
- *Sorting and Ranking.* The model may misinterpret ordering requirements such as ordinal positions, top- k selection, newest-versus-oldest ordering, or rankings conditioned on multiple attributes, leading it to select the wrong item or process items in the wrong order.
- *Quantity and Multi-Target Completeness.* Tasks that specify an exact quantity or require actions over multiple related targets are prone to partial completion. The model may process too few or too many objects, or omit one component of a compound objective.
- *Premature Completion.* The model may declare success after preparing the intended result but before executing the final state-changing action, such as saving an edit, sending a message, submitting a form, or publishing content.
- **Repetitive Action Loops.** The model may repeatedly execute the same action or a short action sequence without making task-relevant progress, often because it fails to recognize that the interface state has not changed or cannot identify a recovery action.
- **Long-Tail Action Selection Failures.** The model may fail to invoke infrequent but critical actions, such as `open`, `ask_user`, or `long_press`, and instead fall back on unsuitable frequent actions such as `click`.

The shared structure of these failures makes them reusable targets for action-level optimization across applications. Their importance, however, is not reflected by their frequency in naturally collected trajectories: rare but consequential patterns can be overwhelmed by more common actions.

Targeted Error-Pattern Data Construction. To correct this coverage imbalance, we combine historical-trajectory mining with active environment interaction. From existing trajectories, we localize the actions that initiate each failure; for repetitive behavior, a visual judge distinguishes unproductive loops from purposeful exploration. For patterns with insufficient historical coverage, agents actively explore executable environments, identify relevant interface structures, construct tasks around them, and collect new rollouts. The resulting corpus preserves realistic failure contexts from the current policy while deliberately increasing the representation of rare but important errors.

Action-Aware Reward and Training. Using this targeted corpus, action RL assigns a structured reward to each predicted action at step t :

$$r_t = F_t (w_{\text{type}} C_t + w_{\text{arg}} C_t Q_t - \lambda_{\text{sens}} S_t - \lambda_{\text{rep}} L_t). \quad (5)$$

where $F_t \in \{0, 1\}$ indicates format validity, C_t measures action-type correctness, and Q_t measures argument quality using pixel distance, lexical similarity, tag matching, or LLM-based judgment. The terms S_t and L_t penalize incorrect sensitive actions and repetitions in the action history, respectively. During training, we observe declining token entropy and progressively shorter reasoning traces. We therefore apply entropy regularization together with lower and upper bounds on reasoning length to prevent policy collapse while avoiding unnecessary verbosity.

2.4.3 Online RL: Learning Long-Horizon Decision Making

Long-horizon GUI tasks are not simply collections of locally correct actions. An action may appear reasonable in the current state yet lead the agent into a future state from which the task becomes difficult or impossible to complete. Action RL improves local decision reliability, but does not fully capture such delayed trajectory-level consequences; we therefore introduce Online RL to optimize end-to-end task success through environment interaction.

Unified Environment Infrastructure for Scalable Rollouts. Learning from complete interactions requires large-scale and reliable rollout generation across heterogeneous GUI environments. Building on the sandbox infrastructure described in Section 2.2.1, we develop a unified environment infrastructure that manages environment allocation, reset, interaction, and evaluation across domains. On Alibaba Cloud, this infrastructure deploys sandbox environments for mobile, computer use, web, and DeepSearch, supporting up to 10,000 concurrent rollouts. Each domain retains its native applications, action spaces, and verification mechanisms, while the shared infrastructure exposes a common environment lifecycle and rollout interface to the training pipeline. For tasks that must be executed on physical phones, we use the virtual-screen mechanism described in Section 2.2.2, allowing multiple application sessions to run concurrently on separate displays of a single device. Across our real-device cluster, this mechanism increases aggregate rollout throughput by approximately $20\times$ compared with running the same cluster without virtual screens. Together, this infrastructure provides scalable rollout capacity across both sandbox and real-device environments for Online RL.

Task-Verifier Synthesis. Building on this rollout infrastructure, we automatically construct executable task-verifier pairs that provide reliable outcome supervision through three phases:

- **Environment State Synthesis.** Coding agents analyze the codebases underlying each sandbox and distill reusable data-injection skills for their applications. Using these skills, the agents construct coherent initial environment states spanning multiple applications. For example, an environment state representing a computer science PhD student preparing a paper submission may contain mutually consistent emails, photos, files, calendar events, and social-media records across the student’s phone. Environment state synthesis is critical because it largely determines the difficulty and diversity of downstream tasks. Generating tasks directly from benchmark-provided states confines task synthesis to the application states and data already available in the benchmark.

- **Task Synthesis.** Conditioned on each synthesized environment state, an LLM generates multiple tasks at different difficulty levels, including workflows spanning multiple applications. An independent LLM judge then filters the generated tasks for feasibility and consistency with the initialized environment state.
- **Verifier Synthesis.** Using the unified environment infrastructure, coding agents autonomously launch corresponding sandboxes, inject synthesized environment states, and construct task-specific code-based verifiers. Each verifier is validated through rollouts from multiple agent models, while a VLM-based judge assesses whether verifier outputs are consistent with observed rollout evidence.

This automated synthesis and execution-validation pipeline produces approximately 10,000 validated task-verifier pairs for Online RL.

Online RL with a Model-Adaptive Task Curriculum. Given the validated task-verifier pairs, we use verifier-guided Online RL to improve long-horizon decision making. Following MAI-UI (Zhou et al., 2025a), we adopt a tailored variant of Group Relative Policy Optimization (GRPO) for GUI interaction trajectories. For each task, the current policy samples a group of K complete interaction trajectories $\{\tau_i\}_{i=1}^K$. The verifier evaluates the final environment state $s_T^{(i)}$ of each trajectory. We compute its binary outcome reward and group-relative advantage as

$$r_i = v_x(s_{\text{final}}^{(i)}) \in \{0, 1\}, \quad \bar{r}_x = \frac{1}{K} \sum_{j=1}^K r_j, \quad \hat{A}_i = \frac{r_i - \bar{r}_x}{\text{Std}(r_1, \dots, r_K) + \epsilon}. \quad (6)$$

where the mean and standard deviation are computed over the K rollouts sampled for the same task x . GRPO updates the policy using these relative advantages. Unlike action RL, which provides localized action-level feedback, this stage optimizes complete trajectories using terminal verifier outcomes, targeting long-horizon planning, state tracking, and intermediate decision making.

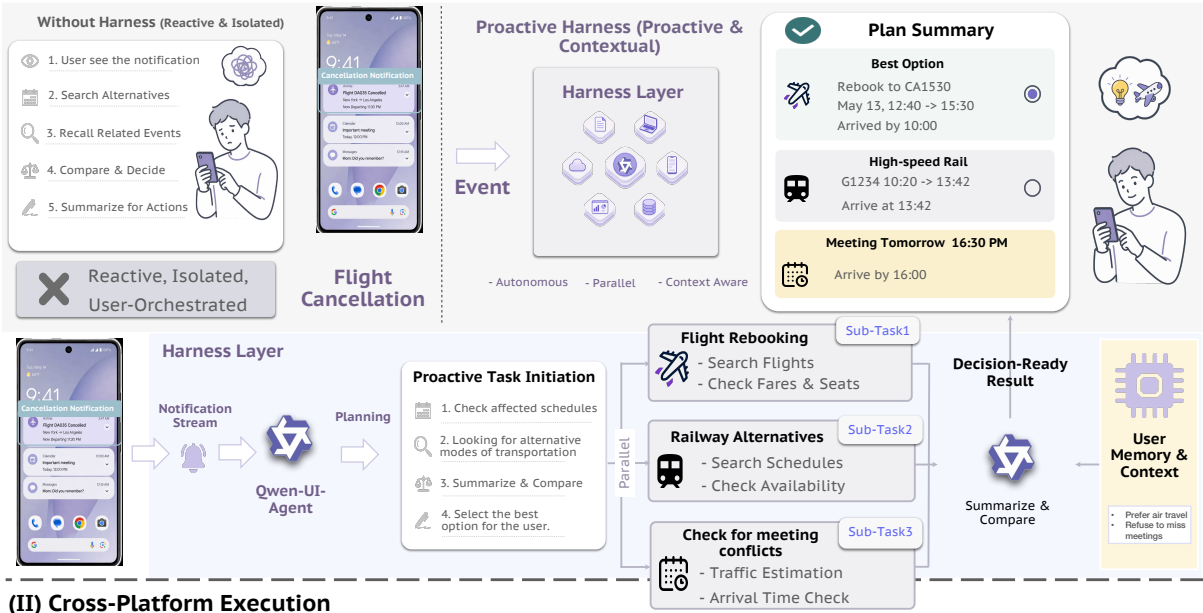
The training value of a task is not fixed: a task that is initially too difficult may become learnable as the policy acquires related capabilities from other tasks. Under the group-relative objective, tasks whose rollouts all fail or all succeed produce no reward variation and therefore provide little immediate learning signal. However, permanently removing difficult tasks would miss the point at which they enter the policy’s learning frontier. We therefore treat task difficulty as a dynamic property of the current policy and organize the curriculum into an active training pool and a monitoring pool. Tasks with intermediate empirical success rates enter the active pool and receive the full rollout budget, while currently unsolvable tasks remain in the monitoring pool with a smaller budget. Once a difficult task begins to produce successful rollouts, it is promoted to the active pool for full-scale training. Mastered tasks are also monitored with a small budget and reactivated if their performance declines. This closed-loop curriculum continually identifies newly learnable tasks while concentrating rollout computation where it provides the most informative relative rewards.

2.5 Harness Layer: Proactive Service and Cross-Platform Task Execution

Strong task execution is necessary, but it is not sufficient for GUI agents to move beyond isolated task execution and become effective assistants in users’ everyday digital lives. Existing agents are largely reactive: users must recognize a need, formulate an instruction, and invoke the agent, even when that need is already evident from signals in their digital environment. Workflows may also span devices, with relevant information appearing on a phone while subsequent actions must be completed on a computer. This creates two fundamental gaps: *when* should an agent initiate assistance, and *how* can it carry the resulting workflow across devices without losing context? We address both through a shared harness layer above the underlying GUI agents. The harness connects signals from the user’s digital environment, maintains a unified representation of ongoing activities, and uses this context to initiate and coordinate tasks across mobile and computer environments.

(I) Proactive Service

From reactive execution -> Proactive service initiation



(II) Cross-Platform Execution

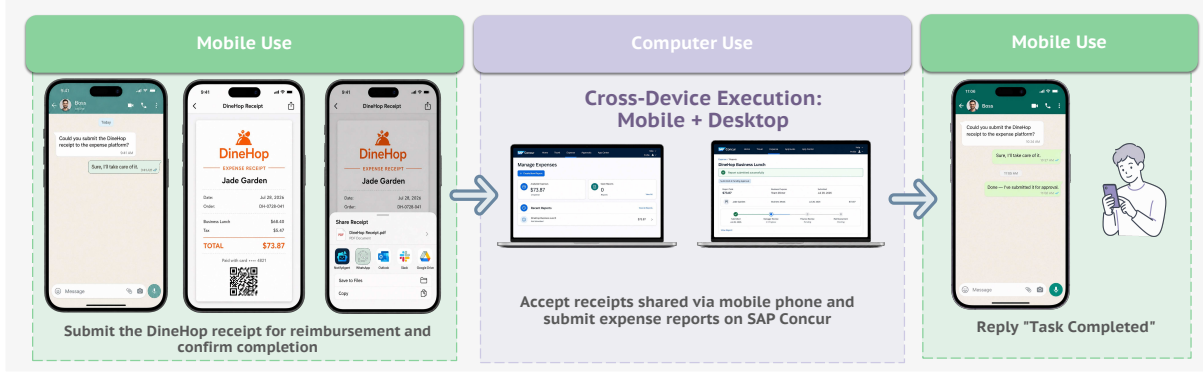


Figure 6: Overview of our harness for proactive service initiation and cross-platform execution. (I) Without the harness, users must manually interpret notifications, retrieve related context, and provide instructions for the agent. The proactive harness instead detects a flight cancellation, reasons over relevant affairs, proactively implement flight rebooking, railway alternatives, and meeting-conflict checks, and report an actionable plan. (II) The harness also maintains task state and support cross-platform execution. In the restaurant-selection example, Qwen-UI-Agent discovers candidates on mobile, organizes them in a desktop spreadsheet, sends the artifact for approval, and saves the selected locations back to mobile.

Built on this layer, we develop two capabilities. *Proactive service* addresses the first gap by shifting task initiation from explicit instructions to to actionable events inferred from the user’s digital signals, enabling the agent to identify and prepare useful assistance before the user asks. *Cross-platform task execution* addresses the second by carrying workflows across phones and computers while preserving shared context, dependencies, and task state. Together, these capabilities extend GUI agents along two complementary dimensions: *when* a task can begin and *where* it can be carried forward.

2.5.1 Proactive Service Based on Mobile Notifications

Mobile Notifications as a High-Value Signal Source. A proactive agent needs signals that are timely, broadly available, and controllable by the user. We use mobile notifications as this interface. A phone aggregates signals from communication, travel, finance, logistics, and scheduling, and notifications expose many events at the moment they become actionable, such as a flight cancellation, a payment deadline, or a schedule change. Compared with continuous screen, audio, or video monitoring, notification access is

lightweight and governed by explicit per-application permissions. Notifications are nevertheless partial and noisy: they report only what an application chooses to surface. The harness therefore treats them as observations rather than complete world state, and gathers additional information only when needed.

From a Notification to a Proactive Workflow. Consider a late-night notification that the user’s flight the following morning has been canceled. A conventional GUI agent would wait for an explicit instruction before searching for alternatives. We instead envision a proactive workflow that goes beyond a simple search. A flight cancellation is rarely an isolated event: it may jeopardize an important meeting, require changes to a hotel reservation, and force a choice between rebooking and rail travel. The key question is whether an agent can anticipate and prepare these interdependent follow-up actions before the user asks. Our harness does so by reconstructing the broader context before requesting the users attention. It parses the cancellation into a structured event, associates it with the ongoing trip and the relevant meeting deadline, retrieves user preferences from memory, queries feasible alternatives, evaluates them against the user’s underlying objective, and presents a decision-ready proposal. Actions with external consequences, such as rebooking or modifying a reservation, remain subject to user confirmation. Figure 11 and Figure 28 of Appendix A.3 show two representative examples of this behavior: a scheduled morning brief that combines the user’s commitments with live weather and commute conditions, and a proactive travel-recovery plan triggered by a flight-cancellation notification.

The harness implements this workflow as a stateful pipeline that parses notifications into structured events, associates them with persistent affairs, derives executable tasks, and uses execution outcomes to update memory and future behavior. Its central abstraction is the *affair*, a persistent representation of an ongoing real-world matter. An *event* records what happened at a particular time; an *affair* captures what is still unfolding across events, applications, and days; and a *task* specifies what should be done next. This separation allows the harness to reason about the user’s underlying goals.

- **Event Perception.** Incoming notifications are processed in short windows. The event parser removes low-value noise and converts useful notifications into structured events containing their source, entities, timestamp, urgency, and available actions. These events are stored with their temporal order and provenance, providing an evidence base for subsequent reasoning.
- **Affair State and Memory.** For each new event, the harness retrieves relevant recent events, active affairs, the current context state, and profile memory. It then associates the event with an existing affair or creates a new one, updating the affair’s lifecycle, key entities, deadlines, unresolved gaps, and supporting evidence. Affairs retain the evolving state of individual trips, purchases, meetings, or family matters, while profile memory stores more stable knowledge such as user preferences, relationships, and behavioral patterns. Together, these two forms of memory combine situation-specific reasoning with long-term personalization.
- **Affair-level Reasoning and Task Formation.** The harness reasons over updated affairs to identify expected next steps, missing information, unfinished actions, time pressure, and dependencies or conflicts across affairs. Conclusions are grounded in observed events or stored user information and are associated with confidence estimates. The task generator then evaluates whether intervening now would provide sufficient value, considering factors such as urgency, consequence, evidential support, and the amount of user effort saved. Depending on this assessment, a task is proposed immediately, placed in a passive to-do list, or suppressed.
- **Proactive Preparation and Controlled Execution.** For a selected task, the harness chooses an appropriate executor, such as an information-seeking tool, a mobile GUI agent, or a computer-use agent. It may perform low-risk preparatory actions before interrupting the user, such as retrieving alternative trains, comparing travel times, or checking refund status. The resulting task card therefore presents concrete, decision-ready options rather than another reminder. Operations with external consequences, including payments, bookings, and sending messages, remain subject to user confirmation.
- **Personalization and Continual Evolution.** The harness continuously refines both its understanding

of the user and its proactive behavior. Profile memory is built from explicit user input, evidence-supported observations, allowing the system to learn preferences, relationships, and routines. Feedback memory records how users respond to proposed tasks, including approval, modification, rejection, ignoring, and completion. These signals reinforce or revise profile entries and calibrate future task value, timing, and degree of automation. Evolution therefore occurs at two levels: profile memory learns *what* matters to the user, while feedback memory learns *when and how* the agent should intervene, making proactive assistance increasingly personalized and useful.

2.5.2 Cross-Platform Task Execution

Users’ digital activities are distributed primarily across mobile devices and computers. As a result, many real-world workflows span multiple platforms whose interfaces, capabilities, and states differ. For example, a user may ask the agent to identify receipts in the mobile photo gallery, transfer the selected images to a desktop, rename them according to their recognized dates, and generate a structured expense spreadsheet. This workflow combines visual selection on the phone, cross-device artifact transfer, and file processing on the computer. Treating each device as an independent episode would fragment the task state, obscure dependencies and failures, and leave the user to coordinate the handoffs. Figure 12 and Figure 29 of Appendix A.3 show two such workflows executed by our harness: the receipt-organization task above, and a parallel restaurant search that operates multiple mobile apps concurrently before summarizing the results on the computer; in both cases, mobile subtasks run on virtual screens without blocking the user’s own use of the device. We therefore implement the harness as a hierarchical planner-executor system with device-addressed actions, shared execution state, and explicit cross-platform operations:

- **Global Perception and Device Grounding.** The harness first enumerates available devices, applications, displays, and workspaces. Each observation is tagged with its application, device, and display, and screenshots are assembled into a labeled multi-device view. This allows the planner to inspect mobile and desktop states jointly. The shared state stores the instruction, application-device mappings, recent observations, artifacts, tool outputs, and subtask results. Device-tagged updates change only the corresponding platform state while preserving the remaining context.
- **Dependency-aware Planning and Executor Selection.** We implement an OpenClaw-like planner with a persistent session, workspace, and tool registry ([OpenClaw Contributors, 2026](#)). The planner decomposes the objective into dependency-aware subtasks and selects an appropriate executor for each. It invokes Qwen-UI-Agent, exposed as a GUI subagent, for visually grounded application interaction. Search, file, CLI, and API tools handle information seeking, deterministic system operations, and synthesis. Each GUI call specifies its instruction and target application and device. Returned results update the session, allowing subsequent steps to be replanned from the latest state.
- **Parallel Multi-agent Execution.** Independent calls to different applications or environments run concurrently. A unique identifier and observation queue isolate each subtask’s actions and feedback. Its action-observation loop is sequential, while different subtasks progress in parallel. On Android, a display manager assigns application-specific subtasks to separate virtual displays on one physical device. Each display provides independent capture and display-addressed input, enabling multiple Qwen-UI-Agent instances to operate different applications concurrently.
- **Hybrid Execution with Qwen-UI-Agent.** Each GUI subtask uses the hybrid action space in Section 2.1.2. Qwen-UI-Agent uses GUI actions for visually grounded interaction and CLI or API actions for structured operations, file processing, and external services. Desktop CLI actions execute shell commands, while Android actions use ADB shell commands; their output, errors, and exit status return as structured observations. Compatible operations can be batched. Platform adapters route actions to the addressed environment and transfer artifacts between devices.

3 Experiments

3.1 Experimental Setup

Models and implementation details. We evaluate three variants of Qwen-UI-Agent, including 27B, 35B-A3B, and 4B models. The 27B model serves as our primary variant for end-to-end agent evaluation across mobile, computer-use, browser, DeepSearch, and general agentic benchmarks. The models are initialized from their corresponding base checkpoints and trained using the pipeline described in Section 2.4. Unless otherwise specified, we follow the official interaction budget and evaluation configuration of each benchmark.

Benchmarks. We evaluate Qwen-UI-Agent across five groups of benchmarks covering real-device mobile use, computer use, browser use, DeepSearch, GUI grounding, and general and agentic capabilities.

- **Real-device mobile use.** We evaluate end-to-end mobile task execution on MobileWorld-Real and AndroidDaily (Sui et al., 2026). MobileWorld-Real, introduced in Section 3.2, contains complex human-written tasks executed on live Android devices in the Chinese mobile ecosystem. AndroidDaily focuses on high-frequency everyday tasks on real Android devices and provides a complementary evaluation of common mobile-use scenarios. We further evaluate on MobileWorld (Kong et al., 2026), which provides a stable simulated environment for challenging long-horizon, cross-application tasks.
- **Computer use.** We evaluate desktop task execution on OSWorld-Verified (Xie et al., 2024) and OSWorld-v2 (Yuan et al., 2026). OSWorld-Verified evaluates task progress across a broad collection of desktop applications and operating-system workflows. OSWorld-v2 places greater emphasis on longer-horizon tasks and reports both partial progress and binary task completion. These benchmarks evaluate whether the model can combine visual interaction, command-line operations, and batched actions to complete complex computer-use workflows.
- **Web navigation and DeepSearch.** We use WebArena (Zhou et al., 2024) to evaluate end-to-end browser interaction on functional websites, where tasks require the model to track evolving page states and complete multi-step user objectives. We separately evaluate DeepSearch on BrowseComp (Wei et al., 2025) and BrowseComp-ZH (Zhou et al., 2025b), which measure persistent information seeking, cross-source evidence verification, and answer synthesis over the English and Chinese web. Together, these benchmarks distinguish direct interaction with web interfaces from research-intensive tasks that require information aggregation and verification.
- **GUI grounding.** We evaluate element localization on ScreenSpot-Pro (Li et al., 2025a), ScreenSpot-V2 (Wu et al., 2024), MMBench-GUI L2 (Xuehui Wang et al., 2025), OSWorld-G-Refined (Xie et al., 2025), and UI-Vision (Nayak et al., 2025). These benchmarks cover high-resolution professional software, mobile and web interfaces, desktop applications, text and icon targets, spatial reasoning, functional reasoning, and goal-oriented grounding instructions.
- **General and agentic capabilities.** We evaluate whether GUI post-training preserves the general capabilities of the base model using MMMU-Pro (Yue et al., 2025), RealWorldQA (xAI, 2024), CharXiv-RQ (Wang et al., 2024c), MathVision (Wang et al., 2024a), AI2D (Kembhavi et al., 2016), MMLU-Pro (Wang et al., 2024b), and IFEval (Zhou et al., 2023). We further evaluate broader agentic capabilities on Tau2-Bench (Barres et al., 2025), Terminal-Bench 2.0 (Merrill et al., 2026), Claw-Eval (Ye et al., 2026), BFCL-v4 (Patil et al., 2025), SkillsBench (Li et al., 2026b), and QwenClawBench (Qwen Team & Data Team, Alibaba Group, 2026). These evaluations measure capabilities such as multimodal reasoning, instruction following, tool use, terminal operation, multi-turn interaction, and autonomous task execution beyond routine GUI navigation.

Evaluation Protocols. We follow the official task sets, environments, and primary scoring protocols unless otherwise stated. For interactive GUI benchmarks, we use the benchmark-defined task and step limits and report the corresponding success, partial-progress, or binary-completion metrics. For

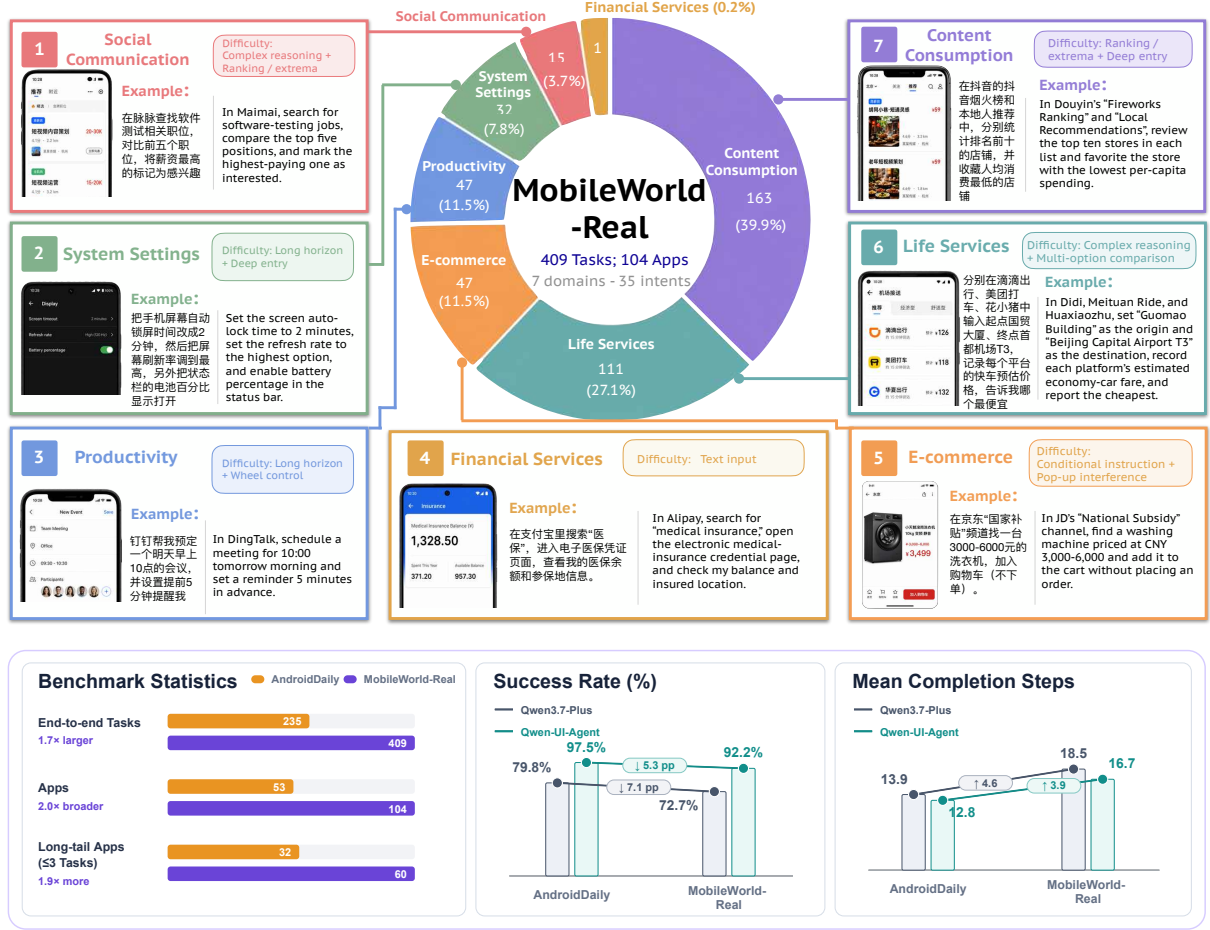


Figure 7: **Overview of MobileWorld-Real, a real-device benchmark with human-written tasks that reflect the breadth of everyday mobile use.** Representative examples and aggregate statistics show broad domain coverage and a long-tailed App distribution. Matched-model results further show lower success rates and longer trajectories than on AndroidDaily, highlighting the challenge of real-world mobile interaction.

MobileWorld-Real, AutoJudge assigns each trajectory one of three outcomes: pass, failed, or env_error. Environment errors are reported separately and excluded from the success-rate denominator, while average trajectory length is computed over successful runs. For WebArena, we manually verify incorrect reference answers and identified errors in the official evaluation scripts before evaluation.

Compared Baselines. We compare Qwen-UI-Agent against three groups of systems: frontier proprietary models, strong open-weight foundation models, and specialized GUI agents. For the general and agentic benchmarks, we additionally compare the 27B model with its base checkpoint to measure capability retention after GUI post-training. We use officially reported results when the evaluation setting is directly comparable and independently reproduce results when a matched evaluation environment is required. The source and evaluation setting of each result are indicated in the corresponding table.

3.2 MobileWorld-Real: Real-world Mobile Use Evaluation Benchmark

Most mobile GUI benchmarks run in sandboxes, where app states can be reset and task outcomes can be verified programmatically (Kong et al., 2026; Rawles et al., 2024). Such environments support controlled comparison, but they can not fully show whether a GUI agent can handle the changing content, account states, and interruptions found on users' phones. We introduce MobileWorld-Real, a real-device benchmark for mobile GUI Agents in the Chinese mobile ecosystem. MobileWorld-Real uses everyday

Table 2: Performance comparison on the GUI-only subset of MobileWorld (Kong et al., 2026), consisting of 117 tasks. We include representative general-purpose VLMs and specialized GUI models. The best result is highlighted in **bold**, and the second-best result is underlined.

MODEL	ACCESS / SIZE	SUCCESS RATE (%)
<i>General-purpose VLMs</i>		
Seed 2.1 Pro (ByteDance Seed, 2026b)	Closed-source	<u>73.2</u>
GPT-5.6 Sol (OpenAI, 2026b)	Closed-source	70.1
Claude Opus 4.8 (Anthropic, 2026)	Closed-source	67.5
Seed 2.0 Pro (ByteDance Seed, 2026a)	Closed-source	63.2
Qwen 3.7 Plus (Qwen Team, 2026a)	397B-A17B	62.3
Gemini 3.1 Pro (Google, 2026a)	Closed-source	58.1
Kimi K2.6 (Moonshot AI, 2026)	1T-A32B	55.6
<i>Specialized GUI Models</i>		
GUI-Owl-1.5-32B-Instruct (Xu et al., 2026)	32B	43.9
MAI-UI-235B-A22B (Zhou et al., 2025a)	235B-A22B	39.7
UI-Venus-1.5-30B-A3B (Gao et al., 2026)	30B-A3B	17.1
<i>Ours</i>		
Qwen-UI-Agent	27B	82.1
Qwen-UI-Agent	35B-A3B	65.0

tasks written by humans and runs them on live Android devices with real Apps, accounts, content, and networks. It evaluates whether an agent can complete the user’s goal while handling the uncertainty and interruptions that arise during real use.

Benchmark Design and Task Coverage. MobileWorld-Real contains 409 end-to-end tasks over 104 apps, each derived from an everyday need identified by a human contributor. These tasks fall into 7 domains of daily mobile use: content consumption, life services, productivity, e-commerce, system settings, financial services, and social communication. Within each domain, the tasks cover multiple user intents. For example, content consumption includes searching for content, browsing feeds, posting, and sharing. To evaluate these tasks under realistic conditions, we execute them on the real-device runtime described in Section 2.2.2, with live account states and dynamically changing online content. This exposes the GUI agent to pop-ups, expired logins, permission requests, CAPTCHAs, and other conditions rarely encountered in sandbox evaluations.

Beyond these environmental challenges, many tasks also involve complex, multi-step workflows. Nearly half are labeled as hard, with common challenges including long-horizon execution, comparison and ranking, reasoning over changing information, navigation to deeply nested app functions, pop-up recovery, and cross-app coordination. The benchmark further captures the ambiguity of real user requests, which do not always provide all the information needed for successful execution. In such cases, the agent must identify the missing information, use `ask_user` to obtain the necessary details, or request user takeover for steps requiring user intervention before continuing the task.

We balance MobileWorld-Real’s realism with the need for repeated evaluation. MobileWorld-Real avoids tasks that depend on a particular past order, cart state, bank card, or other conditions that cannot be set or restored reliably. This ensures high stability of the evaluation results. Figure 7 summarizes MobileWorld-Real’s task coverage, challenge profile, and representative examples.

AutoJudge for Trajectory-level GUI Agent Evaluation. Evaluating task success on live devices is challenging. Unlike sandbox environments, real apps usually do not expose programmatic verifiers for their internal states. Moreover, an unsuccessful run may result from either agent failure or environmental issues, such as CAPTCHAs, expired login sessions, network failures, unavailable services, or infrastructure errors. The final screen alone is often insufficient to distinguish these cases.

Table 3: Performance comparison on real-device mobile benchmarks. MobileWorld-Real is our proposed Chinese real-device mobile GUI benchmark (Section 3.2), and AndroidDaily (Sui et al., 2026) covers high-frequency daily scenarios on real Android devices.

MODEL	ACCESS / SIZE	MOBILEWORLD- REAL	ANDROIDDAILY
<i>Baselines</i>			
Seed 2.1 Pro (ByteDance Seed, 2026b)	Closed-source	88.7	95.2
Gemini 3.1 Pro (Google, 2026a)	Closed-source	86.2	93.8
GPT-5.6 Sol (OpenAI, 2026b)	Closed-source	85.4	92.6
Claude Opus 4.8 (Anthropic, 2026)	Closed-source	84.7	93.0
Qwen 3.7 Plus (Qwen Team, 2026a)	Closed-source	72.7	79.8
Kimi K2.6 (Moonshot AI, 2026)	1T-A32B	62.6	67.6
PhoneBuddy-4B (Tang et al., 2026)	4B	53.5	69.0
UI-Venus-1.5-30B-A3B (Gao et al., 2026)	30B-A3B	33.0	61.7
GUI-Owl-1.5-32B-Instruct (Xu et al., 2026)	32B	32.4	60.9
GELab-Zero-4B-preview (Yan et al., 2025)	4B	31.3	73.4
<i>Ours</i>			
Qwen-UI-Agent	27B	92.2	97.5
Qwen-UI-Agent	35B-A3B	87.4	93.9

To enable scalable and repeatable evaluation, we develop AutoJudge, a trajectory-level evaluator for real-device interaction. The judge prompt specifies explicit decision criteria for assigning each trajectory to one of three outcomes: pass, failed, and env_error. For each run, AutoJudge receives the task instruction and the complete execution trajectory, with every action aligned with its corresponding screenshot. Five independent VLM judges examine the same trajectory, and each produces an outcome label with a short rationale. The final outcome is determined by majority vote. By reporting environment errors separately and excluding them from the success-rate denominator, AutoJudge reduces evaluation noise caused by changing live-device conditions. Appendix A.1 evaluates AutoJudge against independent expert annotations on 666 trajectories, and its exact-match accuracy reaches 92.8%. Figures 22, 23, and 24 present representative examples of the three outcome types.

MobileWorld-Real is designed to complement rather than replace sandbox benchmarks: it measures whether capabilities learned in controlled environments transfer to live devices. All MobileWorld-Real tasks and collected trajectories are held out from training. Since live Apps and account states change over time, we report environment errors separately and retain the complete execution trajectory for auditing and reproducibility.

3.3 Main Results

3.3.1 Mobile-Use Evaluation.

We evaluate Qwen-UI-Agent on the GUI-only subset of MobileWorld (Kong et al., 2026), a challenging and reproducible mobile-use benchmark featuring long-horizon, cross-application workflows across diverse Android applications. The comparison includes both frontier general-purpose VLMs and specialized mobile GUI models from multiple organizations. As shown in Table 2, under the standard 50-step evaluation budget, Qwen-UI-Agent-27B achieves a success rate of 82.1%, establishing a new state of the art. It outperforms GPT 5.6 Sol, Opus 4.8, and Seed 2.1 Pro by 12.0, 14.6, and 8.9 percentage points, respectively. It surpasses the strongest specialized GUI baseline, GUI-Owl-1.5-32B-Instruct at 43.9%, by 38.2 percentage points. Qwen-UI-Agent-35B-A3B also achieves 65.0%, outperforming all specialized GUI baselines and all general-purpose VLMs except Seed 2.1 Pro and Claude Opus 4.8. When the maximum number of steps is increased to 100, the success rates of Qwen-UI-Agent-27B and Qwen-UI-Agent-35B-A3B further rise to 85.5% and 68.4%, respectively. These results demonstrate our strong long-horizon



I'm meeting a friend for coffee at Tianmulu today. Help me search for the detailed address on Amap, open Dianping to find nearby cafés within 1 km, find the most popular one and save its name, then check RedNote and summarize the first three posts to see which coffee people recommend there.

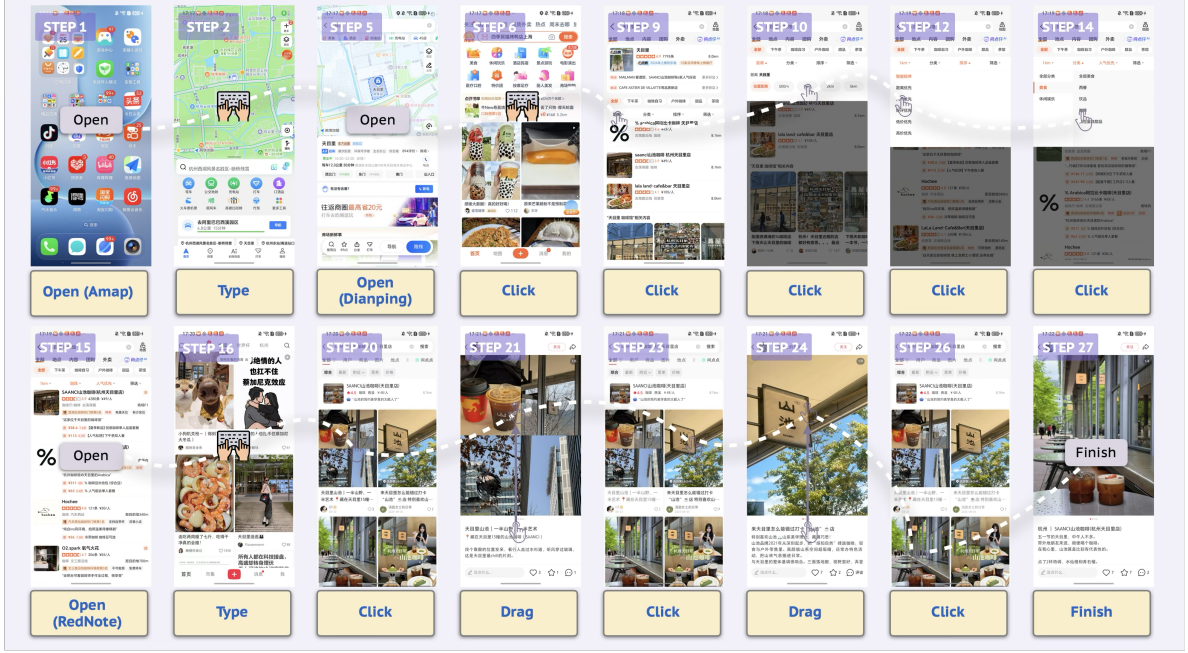


Figure 8: **Demonstration of real-device mobile GUI execution.** The trajectory is rendered as key frames with the executed action annotated beneath each frame. In this cross-app task, the agent looks up the target address on Amap, finds the most popular café nearby on Dianping, and posts a summary of the findings on RedNote.

planning, visual understanding, and reliable execution in complex mobile environments.

We further evaluate Qwen-UI-Agent on two real-device mobile benchmarks: MobileWorld-Real (Section 3.2) and AndroidDaily (Sui et al., 2026). MobileWorld-Real focuses on complex tasks executed on live Android devices in the Chinese mobile ecosystem, while AndroidDaily covers high-frequency daily scenarios on real Android devices.

As shown in Table 3, Qwen-UI-Agent achieves 92.2% on MobileWorld-Real, surpassing all baselines, including the top closed-source models Seed 2.1 Pro at 88.7%, Gemini 3.1 Pro at 86.2%, and GPT 5.6 Sol at 85.4%. This result shows that Qwen-UI-Agent reliably handles long-horizon cross-app execution, deep entry-point navigation within super Apps, and reasoning over changing information on real-world mobile devices. With only 27B parameters, Qwen-UI-Agent offers a favorable balance among task capability, real-device reliability, and deployment efficiency.

On AndroidDaily, Qwen-UI-Agent achieves 97.5%, ranking first among all compared models and outperforming Seed 2.1 Pro at 95.2% and Gemini 3.1 Pro at 93.8%, showing reliable task completion on high-frequency everyday mobile scenarios. The 35B-A3B variant attains 87.4% on MobileWorld-Real and 93.9% on AndroidDaily, while activating only 3B parameters per token, offering substantially higher inference efficiency for deployment.

Qualitative Examples. Figure 8 presents representative real-device execution trajectory of Qwen-UI-Agent, with the executed action annotated beneath each key frame. It highlights autonomous planning in a cross-app scenario: over 27 steps, the agent navigates directly to the deep function entry points of each app, retrieves the address on Amap, identifies the most popular café within 1 km on Dianping, and carries the extracted key information into RedNote as the decision basis for completing the summary.

Table 4: Performance comparison on OSWorld-Verified (Xie et al., 2024).

MODEL	ACCESS / SIZE	SUCCESS RATE (%)
<i>Baselines</i>		
Claude Opus 4.8 (Anthropic, 2026)	Closed-source	83.4
Seed 2.1 Pro (ByteDance Seed, 2026b)	Closed-source	78.8
GPT-5.5 (OpenAI, 2026a)	Closed-source	78.7
Gemini 3.5 Flash (Google, 2026b)	Closed-source	78.4
Gemini 3.1 Pro (Google, 2026a)	Closed-source	76.2
Qwen 3.7 Plus (Qwen Team, 2026a)	Closed-source	73.3
MiniMax M3 (MiniMax, 2026)	428B-A23B	75.2
Kimi K2.6 (Moonshot AI, 2026)	1T-A32B	73.1
GUI-Owl-1.5-32B-Instruct (Xu et al., 2026)	32B	56.5
<i>Ours</i>		
Qwen-UI-Agent	27B	79.5

Table 5: Performance comparison on OSWorld-v2 (Yuan et al., 2026). Partial and Binary denote partial task progress and full task completion, respectively. Lower Steps/task is better.

MODEL	PARTIAL (%)	BINARY (%)	STEPS/TASK	ACTION MODE
<i>Baselines</i>				
Claude Opus 4.8 (Anthropic, 2026)	54.8	20.6	103.0	Batched
GPT-5.5 (OpenAI, 2026a)	49.5	13.0	95.2	Batched
MiniMax M3 (MiniMax, 2026)	22.3	4.6	326.7	Single
Kimi K2.6 (Moonshot AI, 2026)	22.1	4.6	179.3	Single
Qwen 3.7 Plus (Qwen Team, 2026a)	21.5	2.8	173.5	Single
<i>Ours</i>				
Qwen-UI-Agent	40.0	13.9	135.8	Batched

3.3.2 Computer-Use Evaluation

We evaluate the computer-use capability of Qwen-UI-Agent on OSWorld-Verified (Xie et al., 2024) and OSWorld-v2 (Yuan et al., 2026). We report the two benchmarks separately. OSWorld-Verified measures partial progress for 361 computer-use tasks, whereas OSWorld-v2 focuses on longer-horizon workflows and reports both partial progress and binary completion.

As shown in Table 4, Qwen-UI-Agent achieves a success rate of 79.5% on OSWorld-Verified, ranking second among all compared models and trailing only Claude Opus 4.8. It outperforms other leading closed-source foundation models, including Seed 2.1 Pro, GPT-5.5, Gemini 3.5 Flash, and Gemini 3.1 Pro, as well as all evaluated open-weight foundation models. These results demonstrate that Qwen-UI-Agent is competitive with the strongest closed-source models while establishing a substantial advantage over current open-weight models.

As shown in Table 5, Qwen-UI-Agent achieves a partial score of 40.0% and a binary completion rate of 13.9% on OSWorld-v2, placing it competitively among frontier closed-source models. In particular, it exceeds GPT-5.5 by 0.9 percentage points in binary completion, while trailing it by 9.5 points in partial score. Qwen-UI-Agent also substantially outperforms all evaluated open-weight models. Compared with the strongest open-weight baseline, MiniMax M3, Qwen-UI-Agent improves partial progress by 17.7 percentage points and binary completion by 9.3 points. The margins increase to 17.9/9.3 points over Kimi K2.6 and 18.5/11.1 points over Qwen 3.7 Plus. Additionally, Qwen-UI-Agent requires an average of only 135.8 steps per task, substantially fewer than MiniMax M3 and Qwen 3.7 Plus. These results indicate that the hybrid GUI+CLI and batched-action design provides both strong task performance and execution efficiency on long-horizon computer-use workflows.

A key feature of Qwen-UI-Agent is its unified GUI+CLI action space with batched execution. Rather

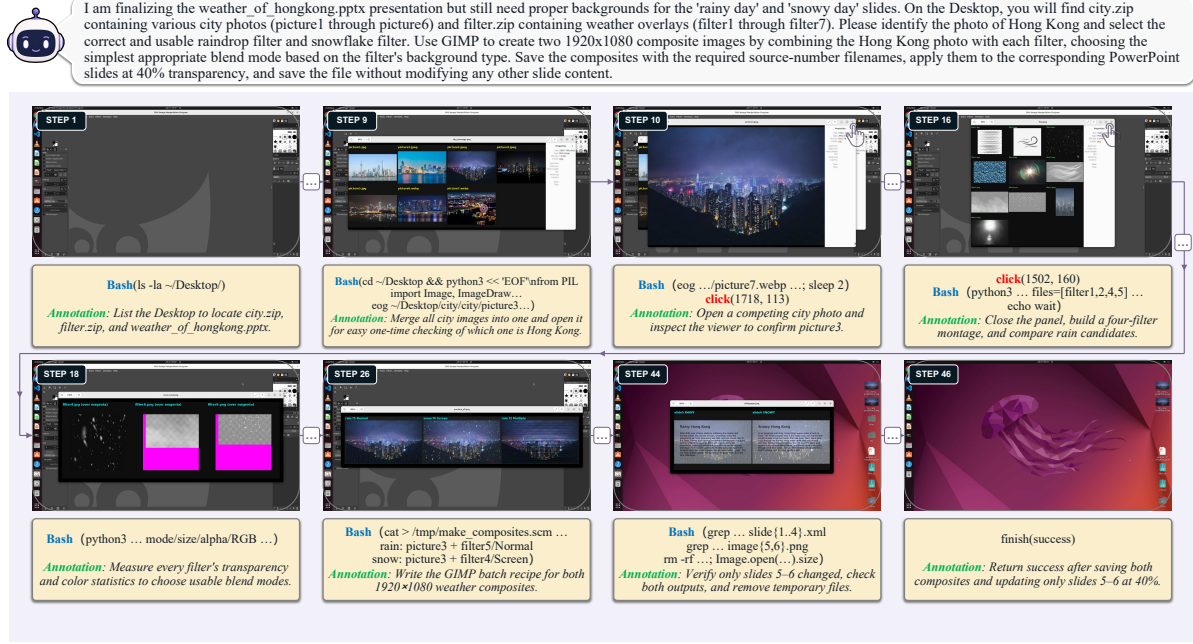


Figure 9: **Demonstration of hybrid GUI+CLI execution in computer-use tasks.** Key frames are shown with the emitted actions, where CLI commands are highlighted in blue, GUI actions in red, and each step carries a short annotation of its intent. In this example, the agent selects the target photo and filters by stitching all candidates into a single montage via CLI for one-shot visual inspection, rather than opening each image individually. It then leverages hybrid execution to combine CLI-based processing with GUI-based verification for the final output.

than treating CLI as an occasional fallback, the model learns to select and interleave GUI and CLI actions within the same trajectory. On OSWorld-Verified and OSWorld-v2, CLI actions constitute **40.7%** and **55.1%** of all actions, respectively, while the corresponding proportions of actions emitted in batched form are **39.6%** and **41.6%**. We provide a detailed analysis of how Qwen-UI-Agent combines GUI and CLI actions in Section 4.2.

Qualitative Examples. Figure 9 and Figure 26 of Appendix A.3 further illustrate the hybrid execution behavior of Qwen-UI-Agent in two representative multi-application workflows, where CLI commands (blue) and GUI actions (red) are highlighted beneath each key frame together with a short annotation of the step intent. In Figure 9, the agent must identify the target city photo and usable weather filters among many candidate images; pure GUI execution would require opening and inspecting every image individually, whereas the agent stitches all candidates into a single montage with one CLI script and verifies them at a single glance, before returning to GUI actions for compositing and presentation editing. Such complementary use of the two interfaces makes hybrid execution substantially more efficient than either GUI or CLI alone, and the agent further batches compatible operations, such as multi-key hotkey sequences or a command followed by a wait, into a single model turn.

3.3.3 Browser-Use and DeepSearch Evaluation

Beyond mobile and computer-use, we evaluate Qwen-UI-Agent on browser-based web navigation and DeepSearch. These capabilities are closely related but assess different aspects of an GUI agent.

Browser Use. We evaluate Qwen-UI-Agent on WebArena (Zhou et al., 2024), which consists of stateful, multi-step tasks over functional websites. During an initial evaluation, we identified multiple incorrect reference answers and errors in the official evaluation scripts. We therefore manually verified the reference

Table 6: Performance comparison on WebArena (Zhou et al., 2024). Baseline results marked with * are obtained from our own evaluation under the same experimental setup.

MODEL	ACCESS / SIZE	SUCCESS RATE (%)
<i>Baselines</i>		
Claude Opus 4.8* (Anthropic, 2026)	Closed-source	71.9
GPT-5.5* (OpenAI, 2026a)	Closed-source	69.5
Gemini 3.1 Pro* (DeepMind, 2025)	Closed-source	65.3
Qwen 3.7 Plus* (Qwen Team, 2026a)	Closed-source	59.0
CUA-GYM-A17B (Wang et al., 2026)	397B-A17B	56.0
Kimi K2.6* (Moonshot AI, 2026)	1T-A32B	55.8
Qwen3.5-397B-A17B (Qwen Team, 2026b)	397B-A17B	54.0
GUI-Owl-1.5-32B-Thinking (Xu et al., 2026)	32B	48.4
Qwen3.5-27B (Qwen Team, 2026b)	27B	41.5
Qwen3.5-35B-A3B (Qwen Team, 2026b)	35B-A3B	40.8
<i>Ours</i>		
Qwen-UI-Agent	27B	73.6
Qwen-UI-Agent	35A3B	69.2
<i>Human Performance (Zhou et al., 2024)</i>		
	–	78.2

Table 7: Performance comparison on DeepSearch benchmarks: BrowseComp (BC) and BrowseComp-ZH (BC-ZH). “–” indicates the result is not reported.

MODEL	ACCESS / SIZE	BC (%)	BC-ZH (%)
<i>Baselines</i>			
GPT-5.5 (OpenAI, 2026a)	Closed-source	90.1	–
Seed 2.1 Pro (ByteDance Seed, 2026b)	Closed-source	86.2	–
Gemini 3.1 Pro (Google, 2026a)	Closed-source	85.9	–
Claude Opus 4.8 (Anthropic, 2026)	Closed-source	84.3	–
UI-TARS-2 (Wang et al., 2025a)	Closed-source	29.6	50.5
Qwen3.5-397B-A17B (Qwen Team, 2026b)	397B-A17B	78.6	70.3
Apodex-1.0-mini (Apodex Team, 2026)	35B-A3B	71.5	80.6
Qwen3.5-27B (Qwen Team, 2026b)	27B	61.0	62.1
GLM-4.7 (Z.ai, 2025)	358B	52.0	66.6
DeepSeek-V3.2 (DeepSeek-AI, 2025)	685B	51.4	65.0
Tongyi-DR-30B (Tongyi DeepResearch, 2025)	30B-A3B	43.4	46.7
<i>Ours</i>			
Qwen-UI-Agent	27B	64.1	75.0

answers and corrected the affected scripts before reporting the final results. Kimi K2.5 likewise reports its WebArena performance using a corrected evaluation setup (Kimi Team, 2026).

As shown in Table 6, the 27B variant of Qwen-UI-Agent achieves a success rate of 73.6%, the highest among all compared models, while the 35A3B variant reaches 69.2%. The 27B variant surpasses the strongest baseline, Claude Opus 4.8 at 71.9%, by 1.7 percentage points, and exceeds GPT-5.5 at 69.5% by 4.1 percentage points. Qwen-UI-Agent also outperforms Gemini 3.1 Pro, Qwen 3.7 Plus, Kimi K2.6, CUA-GYM-A17B, and specialized browser agents such as Claude 3.7 Computer Use and GUI-Owl-1.5-32B-Thinking. Although a 4.6-point gap remains to the reported human performance of 78.2%, the result demonstrates strong end-to-end browser control across heterogeneous web interfaces.

DeepSearch. Simple search is often part of web navigation itself: an agent enters a query into a browser, inspects the ranked results, and opens a potentially relevant page. DeepSearch simplifies this process by replacing many repetitive GUI actions with API-based retrieval and evidence synthesis, allowing the agent to obtain external context more efficiently before continuing the GUI workflow.

Table 8: Performance comparison on GUI grounding benchmarks. ScreenSpot-Pro reports no-zoom results, with zoom-in results shown in parentheses when available. Baseline results marked with * are obtained from our own evaluation.

MODELS	GROUNDING BENCHMARKS				
	SS-PRO(ZOOM IN)	SS-V2	MM-GUI-L2	OSW-G-R	UI-VISION
<i>Baselines</i>					
Qwen 3.7 Plus* (Qwen Team, 2026a)	68.9 (79.0)	96.6	90.5	<u>78.2</u>	<u>68.0</u>
Seed 2.1 Pro* (ByteDance Seed, 2026b)	65.3 (80.7)	96.6	90.9	78.0	62.0
Qwen3.5-4B* (Qwen Team, 2026b)	59.1 (60.3)	94.3	83.4	62.0	32.6
Qwen3.5-35B-A3B* (Qwen Team, 2026b)	64.5 (68.6)	95.2	87.5	63.6	52.3
Qwen3.5-27B* (Qwen Team, 2026b)	68.4 (70.3)	96.1	89.1	67.9	46.6
GUI-Owl-1.5-2B-Instruct (Xu et al., 2026)	57.8 (70.4)	89.7	72.1	62.6	–
GUI-Owl-1.5-4B-Instruct (Xu et al., 2026)	66.8 (75.6)	93.2	83.2	68.4	–
GUI-Owl-1.5-8B-Instruct (Xu et al., 2026)	71.1 (77.8)	93.7	82.5	69.3	–
GUI-Owl-1.5-32B-Instruct (Xu et al., 2026)	72.9 (80.3)	95.3	86.8	69.7	–
UI-Venus-1.5-2B (Gao et al., 2026)	57.7 (64.6)	92.8	80.3	59.4	44.8
UI-Venus-1.5-8B (Gao et al., 2026)	68.4 (73.9)	95.9	88.1	69.7	46.5
UI-Venus-1.5-30B-A3B (Gao et al., 2026)	69.6 (74.8)	96.2	88.6	70.6	54.7
ZoomOnce-4B (Liu et al., 2026)	66.2	95.2	87.6	73.1	40.2
HyMobileAgent-A3B (Shen et al., 2026)	66.5	96.2	89.3	–	–
MAI-UI-2B (Zhou et al., 2025a)	57.4 (62.8)	92.5	82.6	63.5	30.3
MAI-UI-8B (Zhou et al., 2025a)	65.8 (70.9)	95.2	88.8	72.9	40.7
MAI-UI-32B (Zhou et al., 2025a)	67.9 (73.5)	96.5	91.3	75.0	47.1
GTA1-7B (Yang et al., 2025a)	50.1	92.4	78.5	67.7	–
GTA1-32B (Yang et al., 2025a)	63.6	95.2	83.4	72.2	–
UI-Ins-7B (Chen et al., 2025)	52.2	94.0	83.1	–	–
UI-Ins-32B (Chen et al., 2025)	57.0	94.9	84.9	–	–
OpenCUA-7B (Wang et al., 2025b)	50.0	92.3	–	–	29.7
OpenCUA-32B (Wang et al., 2025b)	55.3	93.4	–	–	33.3
OpenCUA-72B (Wang et al., 2025b)	60.8	92.9	–	–	37.3
<i>Ours</i>					
Qwen-UI-Agent-4B	67.8 (74.0)	94.9	87.9	70.5	51.6
Qwen-UI-Agent-35B-A3B	<u>76.1</u> (80.2)	<u>96.7</u>	<u>92.0</u>	74.6	65.9
Qwen-UI-Agent-27B	76.6 (81.5)	97.5	92.6	78.5	70.0

the agent first locates a Douyin video on evidence-based weight loss and extracts its main claims through GUI actions, then calls DeepSearch to verify these claims against research papers, official guidelines, and authoritative sources, since whether and what to comment cannot be decided from the interface alone. The verification conclusions directly determine the subsequent GUI actions, namely posting a comment that corrects the misleading claims and supplements the necessary caveats, forming a closed observe–research–decide–act loop.

3.3.4 GUI Grounding

We evaluate the GUI grounding capabilities of Qwen-UI-Agent across five comprehensive benchmarks: ScreenSpot-Pro, ScreenSpot-V2, MMBench-GUI L2, OSWorld-G-Refined, and UI-Vision. These benchmarks cover high-resolution professional software, general cross-platform interfaces, desktop software environments, and instruction-driven grounding.

As shown in Table 8, the 27B variant achieves strong grounding performance across all five benchmarks, reaching 76.6% on ScreenSpot-Pro without zoom-in, 97.5% on ScreenSpot-V2, 92.6% on MMBench-GUI L2, 78.5% on OSWorld-G-Refined, and 70.0% on UI-Vision. On ScreenSpot-Pro, it ranks first among all compared models both without and with zoom-in, and its score further increases to 81.5% with zoom-in. These results demonstrate accurate element localization across both common interfaces and complex

Table 9: General and agentic performance comparison. Qwen-UI-Agent acquires strong GUI capabilities while preserving the general reasoning and agentic capabilities of its base model. It also substantially outperforms other GUI-specialized models across these benchmarks.

Benchmark	Qwen-UI-Agent	Qwen3.5-72B	UI-Venus-1.5 30B-A3B	GUI-Owl-1.5 32B-Instruct	EvoCUA-32B 20260105	OpenCUA 72B
<i>General capabilities:</i>						
MMMU-Pro	72.4	73.5	32.4	39.5	58.4	31.0
RealWorldQA	83.1	83.1	75.3	76.7	75.4	66.4
CharXiv-RQ	77.7	76.8	44.7	50.9	58.1	39.6
MathVision	82.8	82.0	36.8	50.6	60.8	26.6
AI2D_TEST	91.1	91.9	84.3	84.8	85.7	78.9
MMLU-Pro	86.5	86.0	65.6	73.9	77.1	58.8
IFEval (prompt-level strict)	90.2	90.4	81.3	84.5	64.0	70.6
<i>Agentic capabilities:</i>						
Tau2-Bench	89.9	89.2	22.7	6.1	48.9	14.4
Terminal-Bench 2.0 (Avg 5)	50.1	41.1	3.2	0.0	5.6	9.0
Claw-Eval (Avg 3)	73.5	66.9	30.6	29.6	46.3	26.4
Claw-Eval (Pass@3)	51.8	41.2	5.5	5.5	6.5	0.5
BFCL-v4	74.2	71.3	19.8	32.7	48.8	28.3
SkillsBench (Avg 5)	28.0	24.9	0.5	0.3	3.3	0.0
QwenClawBench (Avg 3)	44.2	48.5	6.4	5.1	18.6	11.4

Evaluation note. All results in this table are independently reproduced in our evaluation environment. The scores may differ from official reports as some evaluation settings are not identical. Unless noted below, we follow the corresponding official benchmark protocol.

- **Tau2-Bench.** We follow the official leaderboard task set, harness, and scoring protocol, but use GPT-5.5 as the user simulator and judge.
- **Terminal-Bench 2.0.** Following the Qwen3.6 evaluation setup, we use the Harbor/Terminus-2 harness with a three-hour timeout and up to 32 CPUs and 48 GB of memory per task. We use temperature 1.0, top- p 0.95, top- k 20, an 80K-token output limit, and a 256K context window, and report the average over five runs.
- **Claw-Eval.** We follow the official 199-task, three-trial protocol and scoring rule, but use GPT-5.5 for both general and multi-turn judging and Qwen3.6-Plus for multi-turn user simulation, replacing the official Gemini 3 Flash/Claude Opus 4.6 setup. We use a sampling temperature of 0.6.
- **SkillsBench.** We follow the Qwen3.6 evaluation setup. We use OpenCode to evaluate 78 self-contained tasks, excluding tasks that depend on external APIs, and report the average over five runs.
- **QwenClawBench.** We follow the official v1.1 task set, three-run protocol, and penalized-hybrid scoring, but replace the default Claude Opus 4.5 judge with GPT-5.2. We also pin the OpenClaw runtime to version 2026.6.1 rather than the moving main image.

professional software, including settings where the target must be inferred from functional or spatial context rather than explicit text.

The grounding capability also remains strong across model scales. On ScreenSpot-Pro, the 35B-A3B and 4B variants achieve 76.1% and 67.8% without zoom-in, with their scores increasing to 80.2% and 74.0% with zoom-in. The 35B-A3B variant further reaches 96.7% on ScreenSpot-V2, 92.0% on MMBench-GUI L2, 74.6% on OSWorld-G-Refined, and 65.9% on UI-Vision, while activating only 3B parameters per token. The corresponding scores of the 4B variant are 94.9%, 87.9%, 70.5%, and 51.6%. Together, these results show that Qwen-UI-Agent provides reliable GUI grounding across common interfaces, professional software, and instruction types, from compact to moderate model scales.

3.3.5 General and Agentic Capabilities

Qwen-UI-Agent develops strong GUI domain knowledge while retaining the general reasoning and agentic capabilities inherited from its base model. As shown in Table 9, it also substantially outperforms specialized GUI models on general and agentic benchmarks. Together, these results indicate that Qwen-UI-Agent remains a broadly capable model rather than becoming a narrow GUI-only model. This broader capability profile may also help it handle out-of-distribution or unusual tasks that require knowledge, reasoning, and tool use beyond routine GUI interaction.



Proactive service: The user's next-morning flight is canceled, putting a 14:00 CTO demo at risk. Before the user asks, the proactive harness searches alternative flights and high-speed trains, evaluates which options can arrive on time, and presents a decision-ready travel recovery plan.

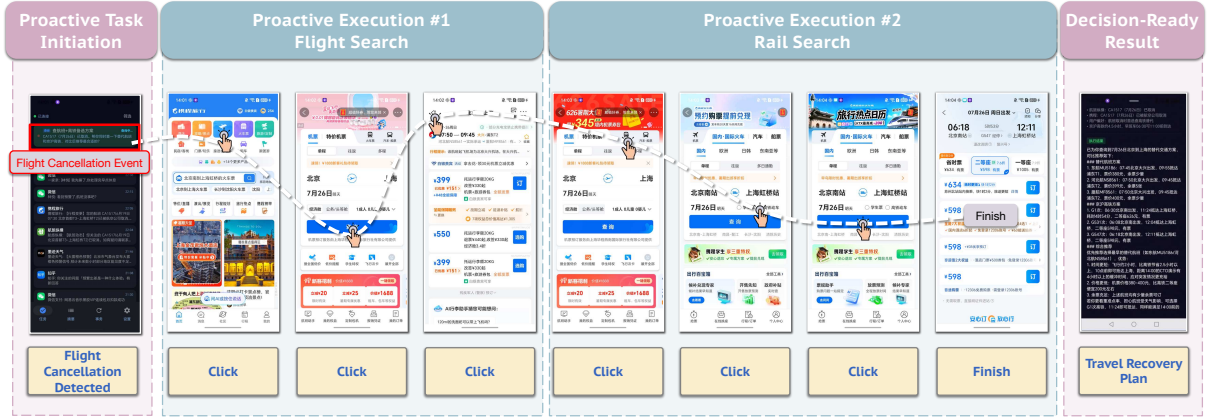


Figure 11: **Demonstration of proactive service based on mobile notifications.** The trajectory is organized into highlighted stages: proactive task initiation from a detected notification, proactive execution phases, and a decision-ready result, with the executed action annotated beneath each key frame. Flight-cancellation recovery: when the user’s next-morning flight is canceled and a 14:00 demo is at risk, the harness proactively searches alternative flights and high-speed trains, evaluates which options arrive on time, and presents a decision-ready travel recovery plan.

3.4 Harness-Enabled Workflows

Beyond benchmark evaluations, we qualitatively examine the workflows enabled by the harness layer introduced in Section 2.5, covering proactive service initiation and cross-platform task execution.

Proactive Service. We present two representative proactive-service trajectories in Figure 11 and Figure 28 of Appendix A.3. Each trajectory is organized into four highlighted stages: proactive task initiation, one or more proactive execution phases, and a decision-ready result. In Figure 11, a scheduled 08:00 trigger combines the user’s morning commitments, extracted from the highlighted notification, with live weather and commute conditions. The harness first retrieves the weather, then reasons over public-transit and taxi options through map navigation, and delivers an actionable morning brief with umbrella advice, departure timing, and key reminders before the user asks. In Figure 28, the harness detects a flight-cancellation notification (highlighted in the first stage) that puts a next-day 14:00 demo at risk. It proactively searches alternative flights and high-speed trains in two execution phases, evaluates which options arrive on time, and presents a decision-ready travel recovery plan, leaving the final booking decision to the user. Both examples show the harness converting passive notifications into contextualized, executable assistance while keeping consequential actions under user control.

Cross-platform Task Execution. Figure 6(II) illustrates a cross-platform workflow for team-building restaurant selection. The harness decomposes the task into four device-addressed stages: Qwen-UI-Agent discovers candidate restaurants on mobile, organizes their ratings, cuisines, and reservation phones into a spreadsheet on the desktop, sends the resulting artifact to the team leader for approval, and finally saves the approved restaurants back to the mobile map application. Throughout the workflow, the shared task state carries the candidate list, the generated spreadsheet, and the approval outcome across devices, so each stage resumes from the latest context rather than restarting from scratch. We present two executed trajectories of this capability in Figure 12 and Figure 29 of Appendix A.3: a receipt-organization workflow that hands mobile gallery selections to GUI+CLI computer use for expense-report generation, and a parallel restaurant search that operates Dianping, Meituan, and Amap concurrently on multiple virtual



Cross-platform service: Search for sushi restaurants across multiple platforms, compare their ratings, and create a local report summarizing the top three options.

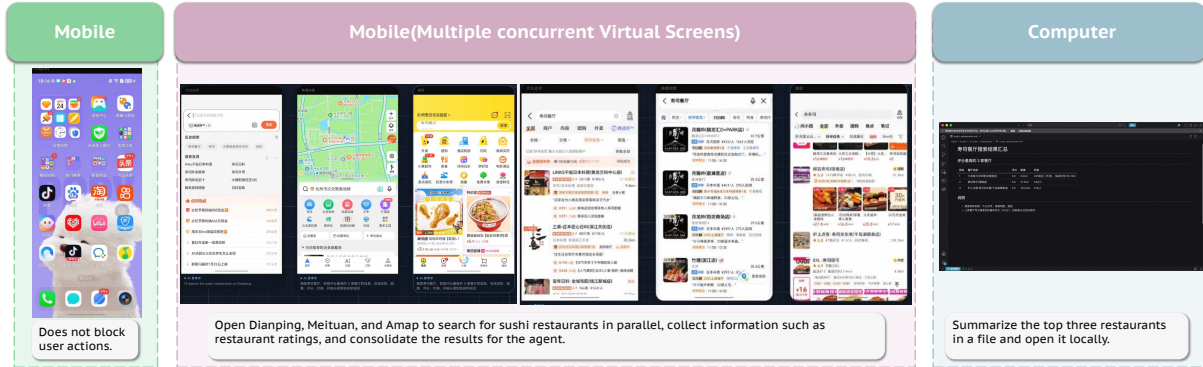


Figure 12: **Demonstration of cross-platform task execution.** In this workflow, mobile subtasks run on virtual screens of the physical device, so execution does not block the user’s own actions. Parallel multi-app search: the agent searches sushi restaurants on Dianping, Meituan, and Amap concurrently through multiple virtual screens, consolidates their ratings, and summarizes the top three options in a local report opened on the computer.

screens; in both cases, mobile subtasks run on virtual screens without blocking the user’s own use of the device. Together, these examples demonstrate that the harness can preserve task state, artifacts, and dependencies as execution moves between mobile and computer environments.

4 Behavioral Analysis of GUI Agent Execution

Aggregate benchmark scores show whether an agent completes a task, but reveal less about how it executes the task, why it fails, and how its behavior changes after RL training. We therefore examine four questions mainly on the execution behavior of Qwen-UI-Agent. (1) Why do models optimized for simulated benchmarks struggle on physical devices? (2) How does Qwen-UI-Agent combine GUI, CLI, and batched actions during task execution, and what distinctive execution patterns emerge from these choices? (3) How does action RL change step-level execution behavior? (4) How does online RL reshape trajectory-level behavior in long-horizon tasks? Together, these analyses complement aggregate metrics by revealing how the agent responds to challenging environments, uses its action space, and changes after RL training.

4.1 How Does a Model’s Limited Real-Device Experience Affect Its Execution Behavior?

Behavioral Analysis of Real-device Failures. Most models have limited direct exposure to physical-device interaction during training. Frontier models such as Qwen 3.7 Plus perform strongly across a broad range of benchmarks, yet their results on MobileWorld-Real suggest that reliable execution on physical devices remains challenging. This raises a more specific behavioral question: when models have limited real-device interaction during training, which execution patterns emerge, and how do these patterns contribute to task failure? To investigate this question, we review every failed Qwen 3.7 Plus trajectory on MobileWorld-Real and AndroidDaily to identify the primary failure modes.

As shown in Table 10, we organize the failures along two dimensions, both rooted in the scarcity of real interaction experience during training: *execution capability limitations* (40.3%), where simulated environments provide little incentive to learn exploration, error recovery, and state tracking, and *real-world scenario challenges* (52.0%), triggered by interface phenomena that sanitized simulators rarely contain. As illustrated in Figure 13, we pair two key frames with the model’s abridged thinking for one representative

Table 10: Failure-pattern distribution over all failed Qwen 3.7 Plus trajectories on real devices.

Dimension	Failure Pattern	Prop.	Typical Behavior
Execution Capability Limitations (40.3%)	Exploration Failure	19.5%	Fail to locate deep in-app entries
	Erroneous Action Loops	14.3%	Repeat ineffective actions
	Lost Execution State	6.5%	Forget finished sub-tasks
Real-world Scenario Challenges (52.0%)	UI Misreading	24.7%	Misread stateful page semantics
	Pop-up Interference	18.2%	Ads, paywalls, CAPTCHAs, blank pages
	Physical Widget Control	9.1%	Overshoot targets, never converge
Others	—	7.7%	Under-execution, premature stop

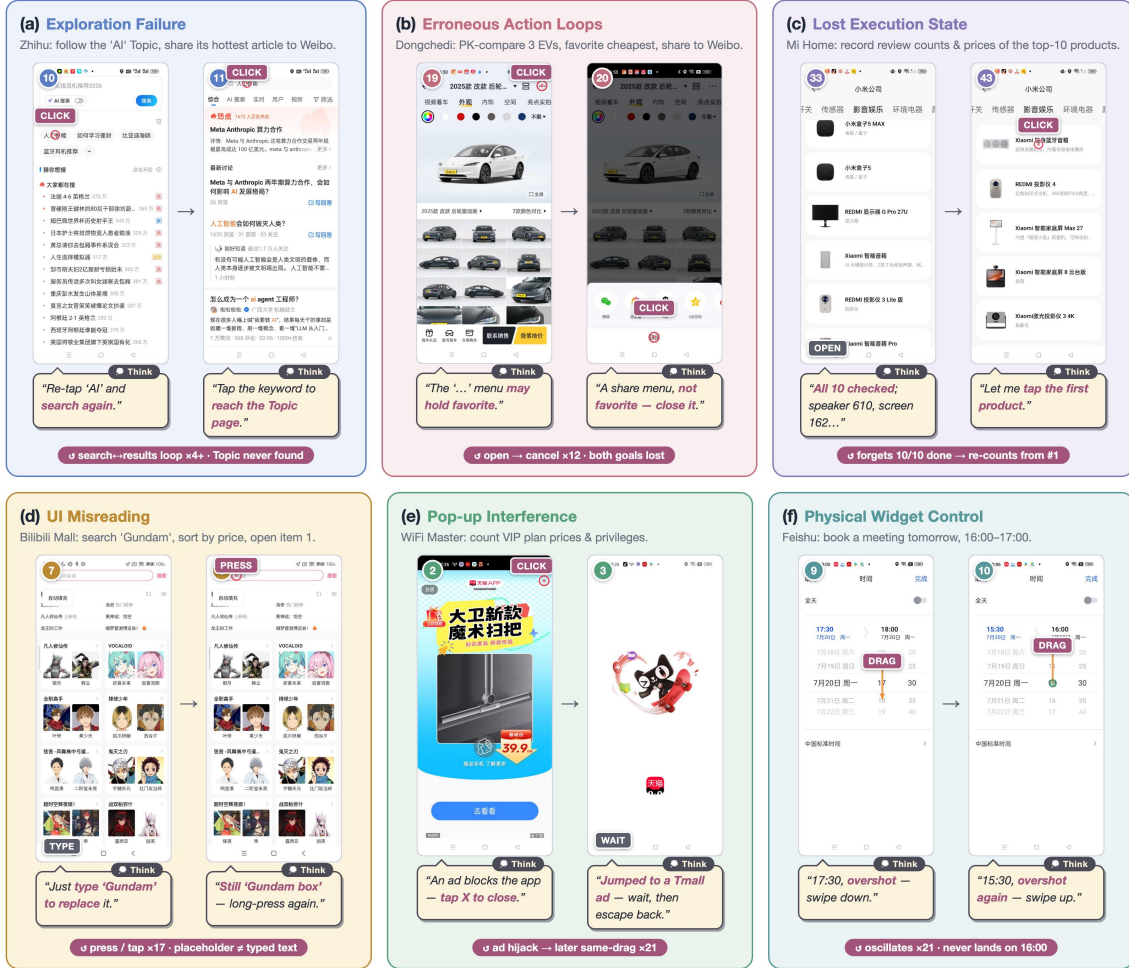


Figure 13: Representative real-device failure patterns of Qwen 3.7 Plus. *Execution capability limitations:* (a) exploration failure, (b) erroneous action loops, (c) lost execution state. *Real-world scenario challenges:* (d) UI misreading, (e) pop-up interference, (f) physical widget control. Speech bubbles quote the model’s abridged thinking, with the critical fallacy in purple.

case of each pattern, and analyze each pattern in detail below.

Execution Capability Limitations. Qwen 3.7 Plus’ failures in this dimension (40.3%) reflect limitations in three execution capabilities: locating less commonly used or deeply nested app functions, recovering from unexpected feedback or interface states, and maintaining task state over long trajectories.

- *Exploration failure* (19.5%) occurs when the target is a less commonly used function hidden behind multiple entry points or deep page hierarchies. In these cases, Qwen 3.7 Plus often revisits a small set

of shallow pages instead of exploring alternative routes (Figure 13(a)). By contrast, Qwen-UI-Agent examines multiple plausible entries before committing to a path.

- *Erroneous action loops* (14.3%) stem from the long tail of real third-party applications, where an action frequently fails to produce the expected transition. Handling such expectation violations is itself a learned skill that simulators, with their highly predictable interaction outcomes, offer few opportunities to acquire. When an action fails to produce the intended result, we find that models with limited exposure to physical-device interaction during training tend to repeat the same action or alternate between two interface states rather than diagnose the failure and revise their plan (Figure 13(b)).
- *Lost execution state* (6.5%) appears in long-horizon tasks that require intermediate information and completed progress to be retained across app switching, interruptions, and multiple subtasks. In these trajectories, Qwen 3.7 Plus may lose track of prior progress and repeat completed subtasks or restart part of the workflow from scratch (Figure 13(c)).

Real-world Scenario Challenges. Failures in this dimension (52.0%) involve interface conditions that are common on physical devices but may be less frequently represented during training: ambiguous page semantics, unexpected interface disruptions, and interactive widgets requiring fine-grained manipulation.

- *UI misreading* (24.7%) arises because real pages carry stateful semantics that pure visual appearance does not disambiguate, such as pre-filled search terms, placeholder hints, and dynamically refreshed lists, whereas simulator interfaces stay clean and static. Models trained in simulated environments consequently misjudge the interface state, most typically mistaking a grey placeholder for user-entered text and repeatedly attempting to clear it instead of typing over it (Figure 13(d)).
- *Pop-up interference* (18.2%) covers advertisements, paywalls, CAPTCHAs, and unresponsive or blank pages, non-deterministic disruptions that simulators exclude by design to keep evaluation reproducible. Lacking any recovery prior, these models either follow the disruption away from the original task or keep retrying the blocked path (Figure 13(e)).
- *Physical widget control* (9.1%) concerns scroll wheels, sliders, and date pickers that demand incremental closed-loop manipulation, where the agent must observe the current value and adjust the swipe magnitude accordingly. In simulated environments such values are typically set by direct text injection or single taps, so the agent applies fixed-magnitude swipes that overshoot the target in both directions and never converge (Figure 13(f)).

The findings above motivate the real-device training and evaluation infrastructure described in Section 2.2.2 and help explain Qwen-UI-Agent’s strong performance on real-world mobile devices.

4.2 A Closer Look at GUI-CLI Coordination and Batched Execution in Qwen-UI-Agent

Qwen-UI-Agent executes computer-use tasks through trajectories that interleave GUI and CLI actions, while selectively emitting multiple actions together in batched form. This hybrid execution process raises two questions: how does the model divide operations between GUI and CLI, and when does it choose to batch multiple actions? We analyze trajectories generated by Qwen-UI-Agent on OSWorld-Verified and OSWorld-v2 and present the resulting behavioral patterns below.

Usage Statistics Across Actions and Tasks. Table 11 quantifies CLI and batched execution at the action and task levels. GUI interaction and CLI operations both serve as primary execution channels in the computer-use trajectories produced by Qwen-UI-Agent. On OSWorld-Verified, CLI operations constitute 40.7% of all actions and appear in 92.0% of tasks. These proportions increase to 55.1% and 98.2% on OSWorld-v2, respectively. The near-balanced distribution of GUI and CLI actions suggests that Qwen-UI-Agent makes flexible and complementary use of both executions. The higher CLI share on OSWorld-v2 partly reflects its task distribution, which contains more workflows that naturally favor programmatic execution, such as merging files, processing structured data, and performing batch file operations. Batched actions account for 39.6% and 41.6% of all actions on OSWorld-Verified and OSWorld-v2, and

Table 11: **GUI+CLI usage and batched-execution statistics on OSWorld-Verified and OSWorld-v2.** Panel (a) reports CLI and batched-action usage at the action and task levels. Panel (b) reports the composition of batched outputs. Mean batch size counts the number of primitive actions in a batch. Differences are computed as OSWorld-v2 minus OSWorld-Verified. Differences are reported in percentage points (pp).

Statistic	Level	OSWorld-Verified	OSWorld-v2	Difference (pp)
<i>(a) Overall CLI and batched-action usage</i>				
CLI	Action	40.7%	55.1%	+14.4
CLI	Task	92.0%	98.2%	+6.2
Batched	Action	39.6%	41.6%	+2.0
Batched	Task	62.1%	88.9%	+26.8
<i>(b) Composition of batched outputs</i>				
GUI-only batches		75.8%	64.7%	-11.1
CLI-only batches		13.1%	15.0%	+1.9
Mixed GUI+CLI batches		11.0%	20.3%	+9.3
Mean primitive actions per batch		3.1	3.1	0.0

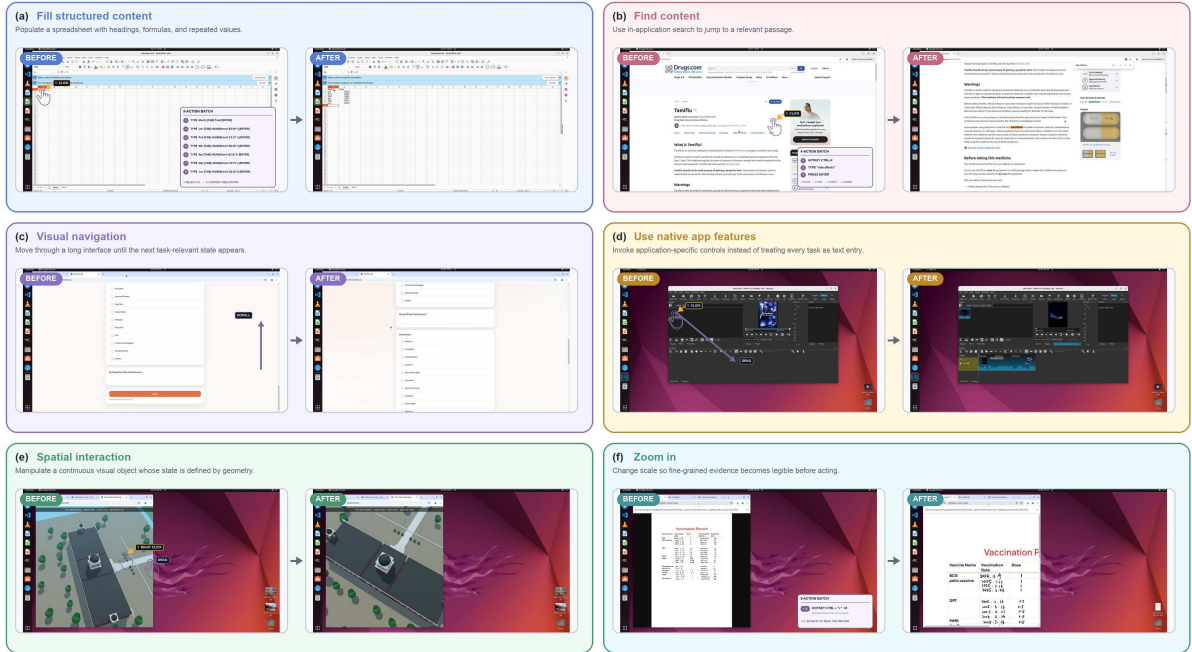


Figure 14: **Representative GUI interaction patterns.** The panels show (a) filling structured spreadsheet content, (b) finding content with in-page search, (c) visual navigation through scrolling, (d) use of a native media-timeline feature, (e) continuous spatial interaction, and (f) zooming in for fine-grained inspection. Each panel pairs the rendered action with its resulting application state.

appear in 62.1% and 88.9% of tasks, respectively, indicating that batched execution is used extensively across both benchmarks. With an average of 3.1 primitive actions per batch, this corresponds to significantly fewer observation–reasoning–execution cycles than issuing each primitive action in a separate model turn, substantially improving execution efficiency. GUI-only batches remain the dominant form, while CLI-only and mixed GUI+CLI batches show that batched execution also supports programmatic operations and coordination between the two execution channels.

When Does Qwen-UI-Agent Use GUI, CLI, and Batched Actions? Beyond aggregate usage statistics, we examine recurring patterns in the trajectories to understand how Qwen-UI-Agent uses GUI interaction, CLI execution, and batched actions in practice. The examples below identify typical scenarios for each



Play the dinosaur jumping game in the browser and achieve a score of at least 100. Do not use browser developer tools or CDP.



Figure 17: **Representative trajectory of tightly coupled GUI-CLI collaboration.** The task requires Qwen-UI-Agent to achieve a score of at least 100 in a browser-based dinosaur game without using DevTools or CDP. GUI observations provide evidence for measuring jump dynamics, diagnosing failed control policies, and recognizing changes between day and night modes, while CLI code iteratively implements stateful key control and adaptive obstacle detection.

dialog, or error. Batched execution therefore groups coherent local transactions rather than merely repeating similar actions.

Case study: Tightly coupled GUI-CLI collaboration in a dynamic task. One particularly interesting trajectory asks Qwen-UI-Agent to play a browser-based dinosaur game and achieve a score of at least 100 without using browser developer tools or the Chrome DevTools Protocol (CDP). Issuing every jump as a separate model action is poorly suited to this task because obstacles continue moving during screenshot acquisition and model inference. Qwen-UI-Agent instead uses CLI to construct and iteratively refine a local controller that operates through rendered screenshots and standard keyboard events. The controller does not access the DOM or internal game state, leaving the visible GUI as its source of task state and the keyboard as its action interface.

The trajectory reveals a tightly coupled pattern of GUI-CLI composition. Successive GUI observations provide the evidence needed to measure jump dynamics, calibrate obstacle-trigger distances, and diagnose errors in jump timing and duration. CLI code converts these observations into explicit perception and control procedures. When repeated trials expose the limitations of stateless, fixed-duration taps, the model introduces explicit action state and holds the jump key until the obstacle has passed. A further challenge arises when the game switches from day to night mode. The resulting contrast inversion invalidates the initial assumption that obstacles correspond to dark pixels. After identifying this change from the rendered frames, Qwen-UI-Agent replaces fixed color thresholding with background-relative segmentation and uses obstacle geometry to handle both ground-level cacti and airborne birds.

The resulting controller survives for 60 seconds while executing 50 jumps and reaches a visible last-run score of 156. This case shows that GUI and CLI provide different but interdependent functions within a single trajectory. The GUI supplies perceptual evidence and execution feedback, while CLI execution turns that evidence into repeatable perception and control procedures that do not require model inference at every game step. By repeatedly moving from GUI observation to CLI-based revision and back to

Table 12: Performance on five error-pattern-specific test sets before and after action RL training.

Model	Confusable-Element Grounding	Sorting and Ranking	Multi-Target Completeness	Premature Completion	Repetitive Action Loops
W/o Action RL	72.8%	76.6%	80.0%	81.0%	72.9%
W/ Action RL	79.1%	80.4%	84.4%	86.2%	82.4%

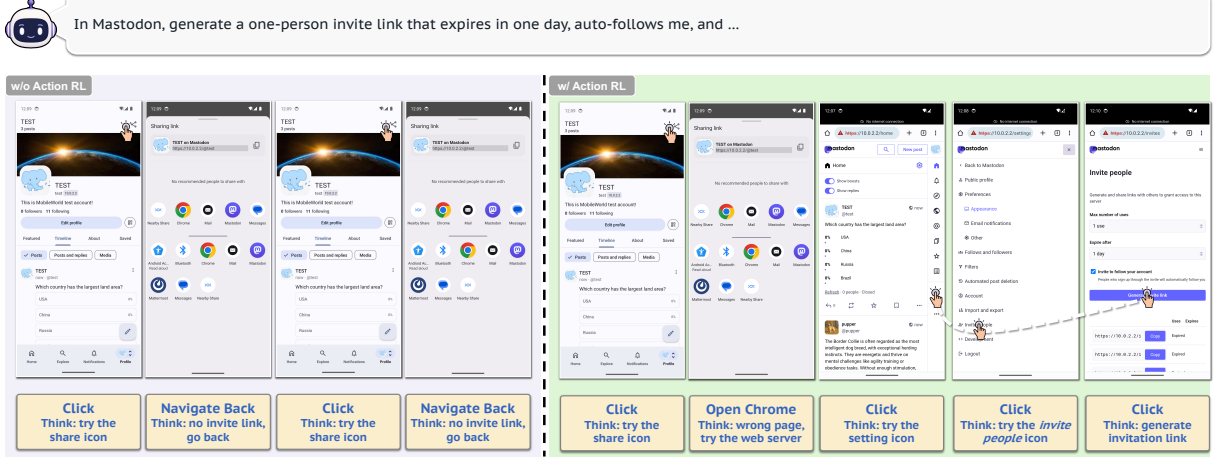


Figure 18: **Action RL case study.** The SFT model navigates to the account page and repeatedly clicks the *share* icon before navigating back, becoming trapped in an ineffective local loop. In contrast, the model after action RL recognizes that it is on the wrong page, switches to exploring Mastodon’s web interface, locates the correct entry point, and successfully configures the target invite link with the required settings.

GUI verification, Qwen-UI-Agent constructs an effective controller without privileged access to the application.

4.3 Step-Level Behavioral Corrections after Action RL

Overall Performance and Behavioral Shift. We further conduct a comparative analysis of the model behaviors before and after action RL training. The primary effect of action RL is not merely an improvement in aggregate task success, but a systematic correction of erroneous action decisions at a fine-grained, step-level granularity. Beyond improving the overall task SR by more than 7%, action RL reduces the number of reasoning tokens by 21.3% while increasing the average number of interaction steps by 8.4%. These changes suggest that the updated policy performs less redundant reasoning and more frequently translates its decisions into environment-grounded actions. More importantly, the performance gains are accompanied by a clear shift in action patterns, characterized by the systematic correction of recurring action errors, which we examine in detail below.

Correction of Recurring Error Patterns. As shown in Table 12, action RL consistently improves performance across all five error-pattern-specific test sets. On grounding tasks that require selecting the correct target among visually or semantically similar candidates, action RL improves SR by 6.3%. It also improves SR on Sorting and Ranking, Quantity and Multi-Target Completeness, and Premature Completion tasks by 3.8%, 4.4%, and 5.2%, respectively. The effect of action RL is particularly evident in correcting short-range action loops. On tasks involving repeated operations over multiple targets or navigation through long scrollable interfaces, action RL improves SR by 9.5%. This gain indicates a stronger ability to recognize when recent actions have failed to produce task-relevant state changes and to switch to an alternative action. Figure 18 presents a representative example in which the agent is asked to generate an invitation link in Mastodon. The model before action RL repeatedly interacts with the profile-sharing icon and eventually reaches the step limit. In contrast, the model after action RL

Table 13: Performance comparison between frequent and long-tail actions.

Action Group	Original Data Proportion	Reward before Action RL	Reward after Action RL
Frequent actions	80.1%	88.3%	92.3%
Long-tail actions	19.9%	71.5%	77.9%

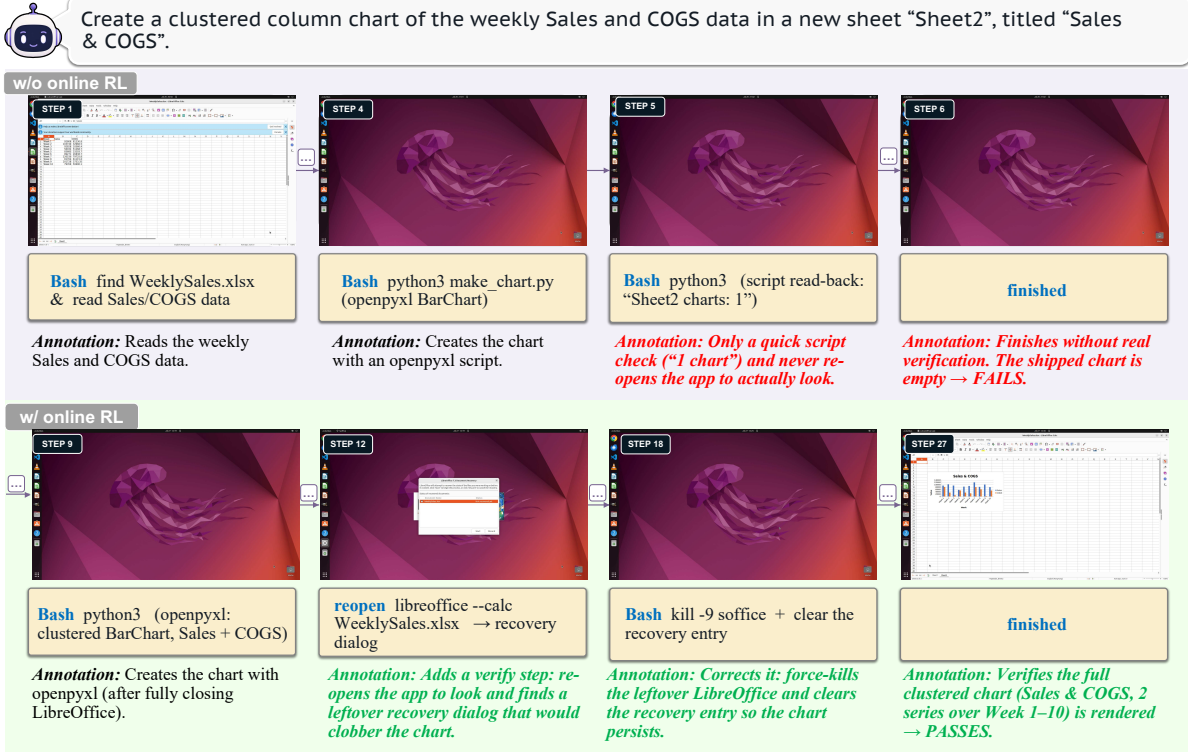


Figure 19: **Online RL elicits verification and self-correction.** On the same task, the online RL policy (bottom) succeeds where the SFT policy (top) fails, because it reopens the application to inspect the actual result and corrects the problem it finds before finishing. The SFT policy trusts a superficial script check that only counts the chart, never looks at the rendered output, and ships an empty chart with no data series, whereas the RL policy reopens the file to verify, finds that a leftover application instance would overwrite the chart, removes it so the chart persists, and confirms that the full clustered chart is actually rendered before terminating.

recognizes that it is on the wrong page, explores an alternative path through Mastodon’s web interface, and reaches the correct interface for configuring the invitation link. This example highlights the central role of step-level action RL: when a short sequence of actions fails to make meaningful progress, the automatically constructed loop-breaking examples teach the policy to revise its immediate action choice rather than persist with an ineffective pattern.

Improved Handling of Long-tail Actions. GUI action trajectories exhibit a highly imbalanced distribution. Frequent actions such as `click`, `drag`, and `type` account for 80.1% of primitive actions, whereas long-tail actions such as `ask_user`, and `long_press` account for only 19.9%. Despite their low frequency, these actions are often critical for reaching specific application states or invoking specialized controls. Pass- N rollouts show that their reward is 16.8% lower than that of frequent actions, reflecting their limited coverage in the training data. Action RL therefore increases the proportion of long-tail action examples to about 40% in the training data, improving their reward by 6.4%. These results demonstrate that action RL can selectively strengthen low-frequency but high-impact action decisions.

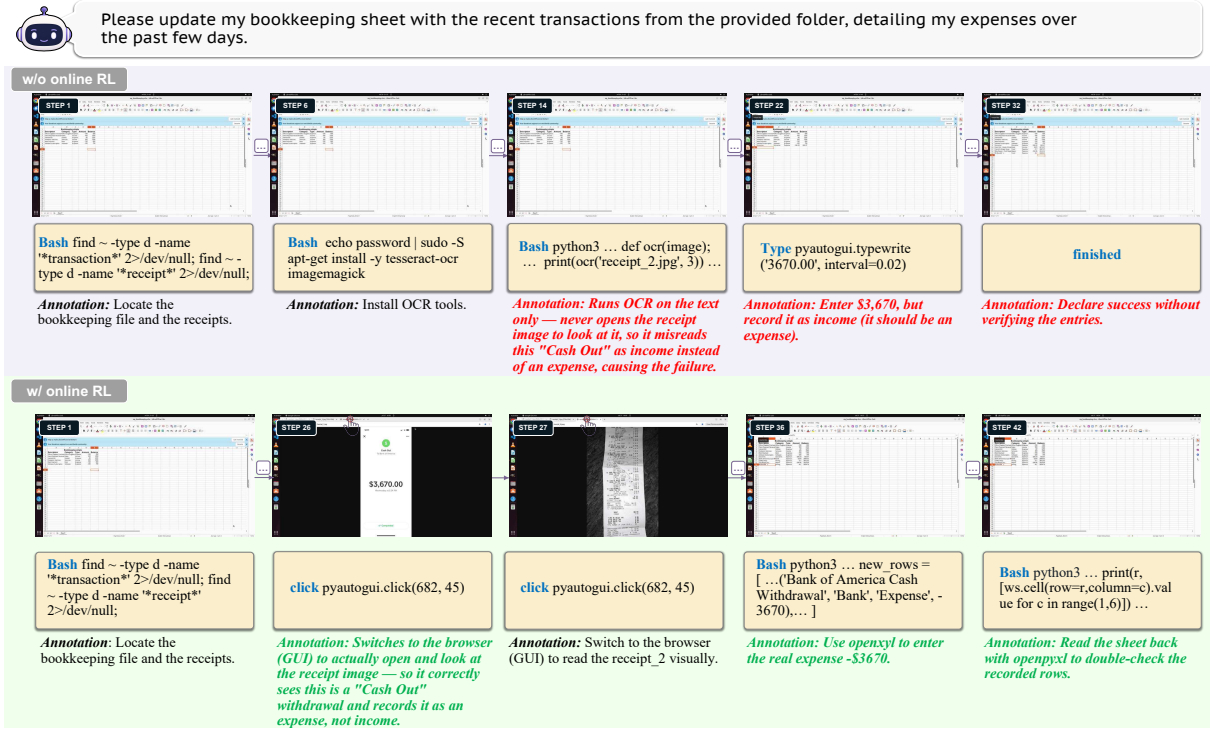


Figure 20: **Online RL elicits emergent cross-modal action.** On the same task, the online RL policy (bottom) succeeds where the SFT policy (top) fails, because it grounds its decisions in what it visually observes and verifies the result before finishing. The SFT policy reads the receipt from OCR text alone, mistakes a Cash Out for income, and declares success without checking, whereas the RL policy opens the receipt image to visually read it, records it correctly as an expense, and reads the spreadsheet back to verify before terminating.

4.4 Trajectory-Level Behavioral Shifts after Online RL

By carefully comparing the trajectories generated before and after online RL across different domains, we identify the emergence of three desirable behavioral patterns.

From Assumed Success to Verified Completion. Before online RL, the policy often terminates immediately after its final state-changing action, assuming that the task has been completed without verifying the intended effect. After online RL, the policy more frequently inspects the resulting state before termination and recovers when verification reveals incomplete execution. Figure 19 illustrates a representative example. Before online RL, the policy performs a sequence of Bash operations and declares completion without checking their effects. After online RL, the policy initially follows the same Bash-based path but adds a verification step, discovers that part of the intended output was not generated, switches to an alternative approach, and ultimately completes the task successfully. Aggregate results on OSWorld show the same trend: the fraction of trajectories containing at least one verification action increases by 14.7%, accompanied by an 11.2% decrease in the false-stop rate—the fraction of self-declared successful episodes that fail the final verifier.

Bash as Hands, GUI as Eyes: Emergent Cross-modal Collaboration. After online RL, the policy shifts from a Bash-leaning strategy to a more balanced use of Bash and GUI actions: the share of GUI actions increases by 6%, while the fraction of trajectories employing both GUI and Bash increases by 10.6 %. Notably, despite the absence of an explicit objective for modality coordination, a useful pattern of cross-modal collaboration emerges: the policy uses Bash to execute state changes efficiently and the GUI to inspect their effects. Consistent with this pattern, the fraction of trajectories exhibiting an



A contemporary Chinese writer rose to fame through a particular novel. One year after the book's second reprint, its television adaptation was aired and received positive reviews. Five years later, the author was selected as one of the centennial figures in a certain industry in China. The title of Chapter 4 in the first volume of the novel contains the name of an animal, and the actor who played one of its villains had previously worked as a bartender. The author was born in the same place as the male lead of a well-known Chinese business drama from the 2000s. In the year when this place was officially established as a city, a major historical event broke out, reshaping China's political and military landscape at the time. What is this place?

	Search (Step #38)	Search (Step #43)	Search (Step #49)	Search (Step #50)	Search Summary
w/o online RL	Condition: Identify a villain actor who previously worked as a bartender. Retrieved Entity: Villain role: Prince Yu; Drama: Nirvana in Fire Search Rationale: This search directly combined the correct drama, villain role, and bartender clue. Condition Met: No. The correct search direction was reached, but the document failed to identify Huang Weide and incorrectly shifted to Andy	Condition: Identify a contemporary writer and the novel that made them famous. Retrieved Entity: Work: Ever Night; Writer: Mao Ni Search Rationale: After confirming Andy On's bartender experience, the document searched his dramas for literary adaptations. Condition Met: No. The candidates did not satisfy the chapter, reprint, selection-year, and birthplace	Condition: Verify the city-establishment year and identify a major event in the same year. Retrieved Entity: City: Guangzhou; City: Establishment year: 1921; Same-year event: Founding of the Communist Party of China Search Rationale: The search tested whether Guangzhou satisfied the administrative and historical-year clues. Condition Met: Yes.	Condition: Verify that the business-production male lead was born in the same city as the writer. Retrieved Entity: Actor: Simon Yam; Actual birthplace: Hong Kong Search Rationale: The actor's birthplace was checked directly to validate the proposed Guangzhou connection. Condition Met: No. The result directly disproved the claim that Simon Yam was born in Guangzhou..	Summary: Summary The place is Guangzhou, China. Guangzhou was officially established as a city in February 1921, becoming China's first incorporated city. The Communist Party of China was founded later that year, an event that profoundly changed China's political and military landscape. The birthplace clue also points to Guangzhou through the male lead of a well-known 2000s business-themed production. Other clues, including the writer's recognition, the animal in Chapter 4's title, and the villain actor's former bartending job, further support this
w/ online RL	Condition: Identify a contemporary writer and the novel that made them famous. Retrieved Entity: Writer: Hai Yan; Novel: Nirvana in Fire Search Rationale: The chapter-title, volume-structure, and television-adaptation clues were used to identify the candidate novel and its author. Condition Met: Yes. Hai Yan was identified as a contemporary writer best known for Nirvana in Fire.	Condition: Identify a villain actor who previously worked as a bartender. Retrieved Entity: Actor: Victor Huang; Drama: Nirvana in Fire Role: Prince Yu Former occupation: Bartender Search Rationale: The actor's early career was searched independently to verify the occupational clue. Condition Met: Yes. His reported early jobs included working as a	Condition: Verify that the business-production male lead was born in the same city as the writer. Retrieved Entity: Business-themed work: Not identified Male lead: Not identified Shared birthplace: Not verified Search Rationale: The search combined the required period, genre, lead role, and Chengdu birthplace. Condition Met: No. No specific work, actor, and reliable birthplace	Condition: Verify the city-establishment year and identify a major event in the same year. Retrieved Entity: City: Chengdu; City: Establishment year: 1928; Same-year event: Northeast Flag Replacement Search Rationale: The search targeted major 1928 events with nationwide political and military consequences. Condition Met: Yes.	Summary: The place is Chengdu, Sichuan, China. Hai Yan became famous for Nirvana in Fire. The novel was reprinted in 2014, and its acclaimed TV adaptation aired in 2015. In 2019, she was named among the "100 Influential Figures of 20 Years of Chinese Online Literature." Chapter 4 of Volume 1 is titled "The Qilin Talent" and Qilin is a mythical animal. Victor Huang, who played the villain Prince Yu, had previously worked as a bartender. Hai Yan was born in Chengdu, which became a city in 1928. The Northeast Flag Replacement also occurred that year, ending Beiyang

Figure 21: **Online RL improves long-horizon search and verification.** Without online RL, the model follows incorrect entity associations and reaches the unsupported answer Guangzhou. With online RL, it identifies the key chain of Hai Yan, *Nirvana in Fire*, Huang Weide, and Chengdu.

execution–verification transition—a state-changing Bash action followed by a read-only GUI inspection—increases from 40.2% to 52.4%. Figure 20 illustrates a representative example. Before online RL, the policy relies only on Bash to execute the task and, without visual information, fails to distinguish income from expenses, introducing errors into the result. After online RL, the policy executes efficiently with Bash and then opens individual receipts in the GUI for visual inspection, allowing it to distinguish the relevant entries and complete the task correctly. This cross-modal loop combines efficient execution with reliable feedback, with Bash serving as the hands and the GUI as the eyes.

Maintaining Constraints Over Long Horizons. Long-horizon tasks often impose additional constraints on the primary goal, including quantities, formats, ranges, and exclusions. Comparing behavior before and after online RL, we find that the policy becomes more reliable at preserving constraints introduced earlier in the instruction throughout execution. Across the two benchmarks, the fraction of tasks satisfying all instruction constraints increases by 8.6% on OSWorld and by 7.5% on BrowseComp-ZH. Figure 21 illustrates a representative BrowseComp-ZH example involving four constraints. In this case, the policy shifts from satisfying only one constraint and reaching an incorrect conclusion to retaining and jointly applying all four, leading to the correct answer.

5 Related Work

Recent work on GUI agents across *mobile-use*, *browser-use*, and *computer-use* increasingly favors native multimodal policies over prompt-only orchestration, with stronger use of self-evolving data engines, online reinforcement learning, and structured memory (Seed, 2025; Zhou et al., 2025a; Gao et al., 2026; He et al., 2024; Wang et al., 2025b; Huang et al., 2026). The field is also moving from benchmark-only optimization toward realistic deployment concerns in terms of real-device execution, privacy, operational reliability, safety, and latency (Cao et al., 2026; Qin et al., 2025; Zhou et al., 2025a; Huang et al., 2026).

Mobile Use. Recent mobile agent works increasingly move beyond screenshot-grounded action predic-

tion toward more capable systems with memory, reflection, RL, and agentic harness. Mobile-Agent-v3.5 (Xu et al., 2026) and UI-Venus-1.5 (Gao et al., 2026) scale unified GUI models across diverse platforms and tasks. Step-GUI (Yan et al., 2025) and UI-TARS-2 (Wang et al., 2025a) investigate data flywheels and multi-turn RL for improving agent capability through iterative interaction. MAI-UI (Zhou et al., 2025a) explores hybrid GUI-MCP operations and device–cloud collaboration. Xiaomi-GUI-0 (Cao et al., 2026) studies real-device closed-loop training that captures state dynamics. Existing mobile agent evaluations mainly focus on three aspects: grounding capability (e.g., ScreenSpot-Pro (Li et al., 2025a)), interactive task completion in controlled environments (e.g., AndroidWorld (Rawles et al., 2024)), and more realistic daily-use scenarios such as MobileWorld (Kong et al., 2026), AndroidDaily (Sui et al., 2026), and RealMobile (Cao et al., 2026).

Browser Use. Compared with mobile agents, browser-use agents can additionally leverage structured information such as DOM trees, accessibility trees, and browser APIs. Early research such as WebVoyager (He et al., 2024) and SeeAct (Zheng et al., 2024) demonstrate the capability of multimodal LLMs to perform end-to-end web navigation by combining visual understanding with browser actions. Recent projects, such as OpenAI Operator (OpenAI, 2025), Google Project Mariner (Google, 2025), and UI-TARS (Seed, 2025), further extend web agents toward autonomous interaction with complex websites and diverse digital environments. Some recent studies like WebWorld (Xiao et al., 2026) and WebEvolver (Fang et al., 2025) also explore world-model-based approaches for web-agent. Web-agent evaluation has evolved from short-horizon webpage navigation toward realistic long-horizon workflows involving multiple websites and complex objectives (Zhou et al., 2024; Xue et al., 2025; Jang et al., 2026).

Computer Use. Computer-use agents (CUA) aim to enable general interaction with desktop environments, covering diverse software and productivity applications. Compared with mobile and web agents, CUAs face a significantly larger action space, more diverse interaction states, and longer horizon, making reliable grounding and execution verification critical. Recent efforts including UI-TARS (Qin et al., 2025), DART-GUI (Li et al., 2025b), OpenCUA (Wang et al., 2025b), UltraCUA (Yang et al., 2025b), EvoCUA (Huang et al., 2026) investigate scalable GUI foundation models by leveraging multimodal pretraining, large-scale interaction data, and agentic reasoning. Recent CUA benchmarks increasingly emphasize robustness, scalability, and real-world deployment challenges. OSWorld-Verified improves evaluation reliability through stricter verification protocols and reduced ambiguity in success assessment (Xie et al., 2024). OSWorld-v2 further scales evaluation toward more complex, long-horizon workflows and diverse real-world computer tasks (Yuan et al., 2026).

6 Conclusion

We presented Qwen-UI-Agent, a real-world centric foundation GUI agent. Guided by a vision of agents that operate on real devices, maintain workflows across platforms, combine GUI interaction with CLI execution, and proactively initiate useful services, Qwen-UI-Agent integrates scalable sandbox environments with a robust real-device mobile runtime, a unified GUI+CLI action space, an agent-driven data flywheel, a training framework combining SFT, action RL, and online RL, and a harness layer for proactive service initiation and cross-platform task execution. Empirically, Qwen-UI-Agent achieves 82.1% on MobileWorld, 92.2% on MobileWorld-Real, and 97.5% on AndroidDaily. It also reaches 79.5% on OSWorld-Verified, and 40.2% on OSWorld-v2. Qwen-UI-Agent further obtains leading or competitive performance across browser use, DeepSearch, and GUI-grounding, while preserving strong general reasoning and agentic capabilities. Together, these results support a systems view of GUI agent development where models, environments, data flywheel, training, and harness are co-designed around real-world usage.

7 Limitations and Future Directions

Limitations. First, we evaluate real-device mobile trajectories using AutoJudge rather than deterministic verifiers or manual expert examination. Deterministic state-based verification is generally infeasible on physical phones because of the restricted access to the internal states of third-party applications. Meanwhile, manually adjudicating trajectories from Qwen-UI-Agent and ten baselines would require substantial expert effort and introduce inter-annotator variation. To ensure a consistent comparison, we apply the same AutoJudge protocol to all systems. On 666 examples independently examined by experts, AutoJudge achieves 92.8% exact-match accuracy. Nevertheless, its remaining errors may introduce minor uncertainty into the reported real-device results. Second, CUA and DeepSearch training at the 35B-A3B scale was still in progress when this technical report was released, so the corresponding results are not included. We plan to update the arXiv version once these results become available. Third, beyond the real-device environments described in this report, higher-fidelity synthetic environments that more closely reproduce real applications provide another path to narrowing the simulation-to-real gap. We have constructed such environments, but training with them was not yet incorporated into the reported model, so this component is not presented in the current report. We plan to open-source the environment synthesis methodology in future work. Fourth, our attempts to fully automate GUI capability development showed that current foundation models cannot yet manage the entire process reliably. Our pipeline is therefore agent-driven rather than fully autonomous and still requires considerable human oversight and intervention to monitor progress and correct errors.

Future Directions. We discuss several directions that we believe warrant further investigation as GUI agents move toward reliable and practical real-world use.

- **Efficient GUI Execution.** GUI agents are capable to complete increasingly complex workflows directly on real devices, marking an encouraging step toward practical use. However, capability alone is not sufficient. Execution efficiency, particularly interaction latency, remains a major obstacle to practical use. Each GUI step typically requires a new observation, model inference, and environment transition, causing latency to accumulate over long trajectories. Virtual-display mechanisms across mobile and desktop systems can partially mitigate this problem by enabling background or parallel execution, but they do not remove the fundamental per-step cost of the perception–reasoning–action loop. Reducing this cost through faster inference, adaptive observation, asynchronous execution, and fewer interaction rounds therefore remains an important for bringing GUI agents into practical real-world use.
- **Harness-Assisted Long-Horizon Workflows.** Our hybrid GUI+CLI action space and support for batched actions improve execution efficiency, while online RL strengthens decision making over long trajectories. A complementary direction is to further exploit the harness as a source of long-horizon support beyond the model itself. Harness-level designs for context compression, task decomposition, memory management, progress tracking, and structured tool use could improve the efficiency and reliability of long-horizon workflows from another perspective.
- **Large-Scale Cross-Domain Online RL.** Long-horizon GUI tasks make online RL costly because rollouts are slow and variable in length, while rewards are often sparse or delayed. Future work should develop more efficient and stable training strategies for these settings, while jointly training across browser-use, mobile-use, and computer-use environments to improve data efficiency and cross-domain generalization (Wang et al., 2025a; Hou et al., 2026).
- **Scalable Synthesis of High-Fidelity Environments.** Real-device environments provide the most faithful interaction dynamics, but remain expensive to scale, reset, and verify. High-fidelity synthetic environments offer a complementary path by approximating real-world application behavior in a form that can be generated and controlled at scale. Scalable synthesis of such environments is important for narrowing the simulation-to-real gap and expanding real-world-oriented training. We have constructed such environments, unfortunately, their integration into model training was not completed in time for

this report. We will open-source the environment synthesis methodology in future work.

- **Safety, User Control, and Personalization.** As GUI agents gain the ability to operate across applications and devices, ensuring that their actions remain safe and under user control becomes increasingly important. In this work, we include safety-sensitive training scenarios that teach the model to invoke the `call_user` action and return control to the user when additional input or authorization is required. More systematic safety evaluations, safety-oriented training objectives, and interpretability-based methods for understanding and constraining agent behavior remain important directions. Personalization is equally important for improving the real-world user experience. Our proactive-service harness provides an initial exploration, but constructing useful user memories and profiles, grounding them in reliable information sources, and giving users control over what information is retained all warrant further investigation.

8 Contributions

Core Contributors

Hanzhang Zhou*, Panrong Tong*, Xu Zhang*, Quyu Kong, Chenglin Cai, Tianyu Xia, Gongjie Zhang, Jianan Zhang, Long Li, Long Chen, Lei Wang, Gaole Dai, Pengxiang Li, Liangyu Chen, Yue Wang[†], Steven Hoi.

Contributors

Hongliang Lu, Miaoyi Zhou, Chen Liu, Yuchen Sun, Yucheng Zhao, Yiran Zhong, Jiahui Zeng, Minggang Wu, Zhixiang Ma, Shaoshun Huang, Shuaihao Zhang.

* Project Co-Leader. [†] Project Leader.

References

- Anthropic. Introducing Claude Opus 4.8. <https://www.anthropic.com/news/claude-opus-4-8>, May 2026. Accessed: 2026-07-19.
- Apodex Team. Apodex-1.0: A verification-centric agent team for discoverative intelligence. Technical report, Apodex, 2026. URL <https://www.apodex.com/blog/apodex-1.0>.
- Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ^2 -Bench: Evaluating conversational agents in a dual-control environment. *arXiv preprint arXiv:2506.07982*, 2025. URL <https://arxiv.org/abs/2506.07982>.
- ByteDance Seed. Seed2.0. <https://seed.bytedance.com/en/seed2>, February 2026a. Released: 2026-02-14.
- ByteDance Seed. Seed2.1 officially released: Advancing AI productivity. <https://seed.bytedance.com/en/blog/seed2-1-officially-released-advancing-ai-productivity>, June 2026b. Accessed: 2026-07-19.
- Wanxia Cao, Chengzhen Duan, Pei Fu, Pengzhi Gao, Niu Lian, Fazhan Liu, Hui Liu, Heng Qu, Qinzhuo Wu, Zhehao Yu, et al. Xiaomi-gui-0 technical report. *arXiv preprint arXiv:2606.31410*, 2026.
- Liangyu Chen, Hanzhang Zhou, Chenglin Cai, Jianan Zhang, Panrong Tong, Quyu Kong, Xu Zhang, Chen Liu, Yuqi Liu, Wenxuan Wang, et al. Ui-ins: Enhancing gui grounding with multi-perspective instruction-as-reasoning. *arXiv preprint arXiv:2510.20286*, 2025.
- Google DeepMind. Gemini 3 pro, 2025. URL <https://deepmind.google/models/gemini/pro/>.
- DeepSeek-AI. DeepSeek-V3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025. URL <https://arxiv.org/abs/2512.02556>.
- Tianqing Fang, Hongming Zhang, Zhisong Zhang, Kaixin Ma, Wenhao Yu, Haitao Mi, and Dong Yu. Webevolver: Enhancing web agent self-improvement with co-evolving world model. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 8970–8986, 2025.
- Changlong Gao, Zhangxuan Gu, Yulin Liu, Xinyu Qiu, Shuheng Shen, Yue Wen, Tianyu Xia, Zhenyu Xu, Zhengwen Zeng, Beitong Zhou, Xingran Zhou, Weizhi Chen, Sunhao Dai, Jingya Dou, Yichen Gong, Yuan Guo, Zhenlin Guo, Feng Li, Qian Li, Jinzhen Lin, Yuqi Zhou, Linchao Zhu, Liang Chen, Zhenyu Guo, Changhua Meng, and Weiqiang Wang. UI-Venus-1.5 Technical Report. *arXiv preprint arXiv:2602.09082*, 2026. URL <https://arxiv.org/abs/2602.09082>.
- Genymobile. scrpy: Display and control your Android device. <https://github.com/Genymobile/scrpy>, 2018. Accessed: 2026-07-28.
- Google. Project mariner. <https://labs.google.com/mariner/landing/>, 2025.
- Google. Gemini 3.1 Pro Preview. <https://ai.google.dev/gemini-api/docs/models/gemini-3.1-pro-preview>, February 2026a. Google AI for Developers; accessed: 2026-07-19.
- Google. Gemini 3.5 Flash. <https://ai.google.dev/gemini-api/docs/models/gemini-3.5-flash>, May 2026b. Google AI for Developers; accessed: 2026-07-19.
- Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan, and Dong Yu. Webvoyager: Building an end-to-end web agent with large multimodal models, 2024. URL <https://arxiv.org/abs/2401.13919>.
- Zhenyu Hou, Yujiang Li, Jie Tang, and Yuxiao Dong. Single-rollout asynchronous optimization for agentic reinforcement learning. *arXiv preprint arXiv:2607.07508*, 2026.

-
- Xuhao Hu, Xi Zhang, Haiyang Xu, Kyle Qiao, Jingyi Yang, Xuanjing Huang, Jing Shao, Ming Yan, and Jieping Ye. ToolCUA: Towards optimal GUI-tool path orchestration for computer use agents. *arXiv preprint arXiv:2605.12481*, 2026. URL <https://arxiv.org/abs/2605.12481>.
- Mianqiu Huang, Taofeng Xue, Chong Peng, Jinrui Ding, Sicheng Fan, Jiale Hong, Yufei Gao, Xiaocheng Zhang, Linsen Guo, Xin Yang, et al. Evocua-1.5: Online reinforcement learning for multi-turn computer-use agents. *arXiv preprint arXiv:2607.09773*, 2026.
- Lawrence Keunho Jang, Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. Odysseys: Benchmarking web agents on realistic long horizon tasks. *arXiv preprint arXiv:2604.24964*, 2026.
- Jina AI. Jina reader: Converting URLs to LLM-friendly input. <https://jina.ai/reader>, 2024. Accessed: 2026-07-28.
- Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European Conference on Computer Vision (ECCV)*, pp. 235–251, 2016. URL <https://arxiv.org/abs/1603.07396>.
- Kimi Team. Kimi K2.5: Visual agentic intelligence. <https://www.kimi.com/blog/kimi-k2-5.html>, 2026. Accessed: 2026-07-21.
- Quyu Kong, Xu Zhang, Zhenyu Yang, Nolan Gao, Chen Liu, Panrong Tong, Chenglin Cai, Hanzhang Zhou, Jianan Zhang, Liangyu Chen, et al. Mobileworld: Benchmarking autonomous mobile agents in agent-user interactive and mcp-augmented environments. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6142–6167, 2026.
- Chenxin Li, Zhengyao Fang, Zhengyang Tang, Pengyuan Lyu, Xingran Zhou, Xin Lai, Fei Tang, Liang Wu, Yiduo Guo, Weinong Wang, Junyi Li, Yi Zhang, Yang Ding, Huawen Shen, Sunqi Fan, Shangpin Peng, Zheng Ruan, Anran Zhang, Benyou Wang, Chengquan Zhang, and Han Hu. PhoneHarness: Harnessing phone-use agents through mixed GUI, CLI, and tool actions. *arXiv preprint arXiv:2606.14832*, 2026a. URL <https://arxiv.org/abs/2606.14832>.
- Kaixin Li, Meng Ziyang, Hongzhan Lin, Ziyang Luo, Yuchen Tian, Jing Ma, Zhiyong Huang, and Tat-Seng Chua. Screenspot-pro: GUI grounding for professional high-resolution computer use. In *Workshop on Reasoning and Planning for Large Language Models*, 2025a. URL <https://openreview.net/forum?id=XaKNDIAHas>.
- Pengxiang Li, Zechen Hu, Zirui Shang, Jingrong Wu, Yang Liu, Hui Liu, Zhi Gao, Chenrui Shi, Bofei Zhang, Zihao Zhang, Xiaochuan Shi, Zedong Yu, Yuwei Wu, Xinxiao Wu, Yunde Jia, Liuyu Xiang, Zhaofeng He, and Qing Li. Efficient multi-turn rl for gui agents via decoupled training and adaptive data curation. *arXiv preprint arXiv:2509.23866*, 2025b.
- Xiangyi Li, Wenbo Chen, Yimin Liu, Shenghan Zheng, Xiaokun Chen, Yifeng He, Yubo Li, Bingran You, Haotian Shen, Jiankai Sun, et al. SkillsBench: Benchmarking how well agent skills work across diverse tasks. *arXiv preprint arXiv:2602.12670*, 2026b. URL <https://arxiv.org/abs/2602.12670>.
- Niu Lian, Alan Chen, Zhehao Yu, Chengzhen Duan, Fazhan Liu, Hui Liu, Pei Fu, Jian Luan, Yaowei Wang, Shu-Tao Xia, and Jinpeng Wang. UI-MOPD: Multi-platform on-policy distillation for continual GUI agent learning. *arXiv preprint arXiv:2607.04425*, 2026. URL <https://arxiv.org/abs/2607.04425>.
- Chen Liu, Ling Chen, Hanzhang Zhou, Liangyu Chen, Chenglin Cai, Xin Yu, Steven Hoi, and Yue Wang. One forward beats two: Innerzoom for accurate and efficient gui grounding, 2026. URL <https://arxiv.org/abs/2606.30084>.

-
- Mike A. Merrill, Alexander G. Shaw, Nicholas Carlini, Boxuan Li, Harsh Raj, Ivan Bercovich, Lin Shi, Jeong Yeon Shin, Thomas Walshe, E. Kelly Buchanan, et al. Terminal-Bench: Benchmarking agents on hard, realistic tasks in command line interfaces. *arXiv preprint arXiv:2601.11868*, 2026. URL <https://arxiv.org/abs/2601.11868>.
- Microsoft. Playwright: Fast and reliable end-to-end testing for modern web apps. <https://playwright.dev>, 2020. Accessed: 2026-07-28.
- MiniMax. MiniMax M3: Frontier coding, 1m context, native multimodality. <https://www.minimax.io/blog/minimax-m3>, June 2026. Accessed: 2026-07-19.
- Moonshot AI. Kimi K2.6: Advancing open-source coding. <https://www.kimi.com/blog/kimi-k2-6>, April 2026. Accessed: 2026-07-19.
- Shravan Nayak, Xiangru Jian, Kevin Qinghong Lin, Juan A. Rodriguez, Montek Kalsi, Rabiul Awal, Nicolas Chapados, M. Tamer Özsu, Aishwarya Agrawal, David Vazquez, Christopher Pal, Perouz Taslakian, Spandana Gella, and Sai Rajeswar. Ui-vision: A desktop-centric gui benchmark for visual perception and interaction, 2025. URL <https://arxiv.org/abs/2503.15661>.
- OpenAI. Introducing operator. <https://openai.com/index/introducing-operator/>, 2025.
- OpenAI. GPT-5.5 system card. <https://deploymentsafety.openai.com/gpt-5-5>, April 2026a. Accessed: 2026-07-19.
- OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. <https://openai.com/index/gpt-5-6/>, July 2026b. Accessed: 2026-07-20.
- OpenClaw Contributors. OpenClaw: Personal ai assistant. GitHub repository, 2026. URL <https://github.com/openclaw/openclaw>. Accessed: 2026-07-20.
- Shishir G. Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. The Berkeley function calling leaderboard (BFCL): From tool use to agentic evaluation of large language models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. BFCL-v4 leaderboard: <https://gorilla.cs.berkeley.edu/leaderboard.html>.
- Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, et al. Ui-tars: Pioneering automated gui interaction with native agents. *arXiv preprint arXiv:2501.12326*, 2025.
- Qwen Team. Qwen3.7-Plus: Multimodal agent intelligence. <https://qwen.ai/blog?id=qwen3.7-plus>, May 2026a. Accessed: 2026-07-19.
- Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026b. URL <https://qwen.ai/blog?id=qwen3.5>.
- Qwen Team and Data Team, Alibaba Group. QwenClawBench: Real-user-distribution benchmark for OpenClaw agents. <https://github.com/SKYLENAGE-AI/QwenClawBench>, April 2026. Accessed: 2026-07-24.
- Christopher Rawles, Sarah Clinckemaillie, Yifan Chang, Jonathan Waltz, Gabrielle Lau, Marybeth Fair, Alice Li, William Bishop, Wei Li, Folawiyo Campbell-Ajala, Daniel Toyama, Robert Berry, Divya Tyamagundlu, Timothy Lillicrap, and Oriana Riva. Androidworld: A dynamic benchmarking environment for autonomous agents, 2024. URL <https://arxiv.org/abs/2405.14573>.
- Remote Android Project. Redroid: A multi-arch, gpu-enabled android in cloud container solution. <https://github.com/remote-android/redroid-doc>, 2024.

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

Serper. Serper: Google search API. <https://serper.dev>, 2023. Accessed: 2026-07-28.

Huawen Shen, Zhengyang Tang, Shangpin Peng, Liang Wu, Anran Zhang, Weinong Wang, Yiduo Guo, Chenxin Li, Zhengyao Fang, Yang Ding, Junyi Li, Fei Tang, Zheng Ruan, Yi Zhang, Xingran Zhou, Dingchen Yang, Sunqi Fan, Zhiyi Wan, Han Hu, Xin Lai, Pengyuan Lyu, and Chengquan Zhang. HyMobileAgent: Data-Environment Co-Scaling for Efficient GUI Agents. *arXiv preprint arXiv:2607.14548*, 2026. URL <https://arxiv.org/abs/2607.14548>.

Yifan Sui, Xin Huang, Hongbing Li, Fang Xu, Jiahe Lv, Haolong Yan, Yeqing Shen, Litao Liu, Zhimin Fan, Ziyang Meng, et al. Androiddaily: A verifiable benchmark for mobile gui agents on real-world closed-source applications. *arXiv preprint arXiv:2605.27761*, 2026.

Zhengyang Tang, Xin Lai, Pengyuan Lyu, Xinyuan Wang, Tianyi Bai, Chenxin Li, Yiduo Guo, Huawen Shen, Yuxuan Liu, Junyi Li, Zhengyao Fang, Yang Ding, Yi Zhang, Weinong Wang, Xingran Zhou, Liang Wu, Fei Tang, Sunqi Fan, Shangpin Peng, Zheng Ruan, Anran Zhang, Benyou Wang, Ji-Rong Wen, Rui Yan, Chengquan Zhang, and Han Hu. Phonebuddy: Training open models for agentic phone use, 2026. URL <https://arxiv.org/abs/2606.23049>.

Tongyi DeepResearch. Tongyi DeepResearch technical report. *arXiv preprint arXiv:2510.24701*, 2025. URL <https://arxiv.org/abs/2510.24701>.

Bowen Wang, Dunjie Lu, Junli Wang, Tianyi Bai, Shixuan Liu, Zhipeng Zhang, Haiquan Wang, Hao Hu, Tianbao Xie, Shuai Bai, Dayiheng Liu, Que Shen, Junyang Lin, and Tao Yu. CUA-GYM: Scaling verifiable training environments and tasks for computer-use agents. *arXiv preprint arXiv:2605.25624*, 2026. URL <https://arxiv.org/abs/2605.25624>.

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

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with MATH-Vision dataset. *arXiv preprint arXiv:2402.14804*, 2024a. URL <https://arxiv.org/abs/2402.14804>.

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

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*, 2024b. URL <https://arxiv.org/abs/2406.01574>.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. CharXiv: Charting gaps in realistic chart understanding in multimodal LLMs. *arXiv preprint arXiv:2406.18521*, 2024c. URL <https://arxiv.org/abs/2406.18521>.

-
- Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. BrowseComp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*, 2025. URL <https://arxiv.org/abs/2504.12516>.
- Zhiyong Wu, Zhenyu Wu, Fangzhi Xu, Yian Wang, Qiushi Sun, Chengyou Jia, Kanzhi Cheng, Zichen Ding, Liheng Chen, Paul Pu Liang, and Yu Qiao. Os-atlas: A foundation action model for generalist gui agents, 2024. URL <https://arxiv.org/abs/2410.23218>.
- xAI. Grok-1.5 vision preview. <https://x.ai/news/grok-1.5v>, 2024. RealWorldQA benchmark released alongside the blog post; dataset at <https://huggingface.co/datasets/xai-org/RealworldQA>.
- Zikai Xiao, Jianhong Tu, Chuhang Zou, Yuxin Zuo, Zhi Li, Peng Wang, Bowen Yu, Fei Huang, Junyang Lin, and Zuozhu Liu. Webworld: A large-scale world model for web agent training. *arXiv preprint arXiv:2602.14721*, 2026.
- Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments, 2024.
- Tianbao Xie, Jiaqi Deng, Xiaochuan Li, Junlin Yang, Haoyuan Wu, Jixuan Chen, Wenjing Hu, Xinyuan Wang, Yuhui Xu, Zekun Wang, Yiheng Xu, Junli Wang, Doyen Sahoo, Tao Yu, and Caiming Xiong. Scaling computer-use grounding via user interface decomposition and synthesis, 2025. URL <https://arxiv.org/abs/2505.13227>.
- Haiyang Xu, Xi Zhang, Haowei Liu, Junyang Wang, Zhaozai Zhu, Shengjie Zhou, Xuhao Hu, Feiyu Gao, Junjie Cao, Zihua Wang, Zhiyuan Chen, Jitong Liao, Qi Zheng, Jiahui Zeng, Ze Xu, Shuai Bai, Junyang Lin, Jingren Zhou, and Ming Yan. Mobile-agent-v3.5: Multi-platform fundamental gui agents. *arXiv preprint arXiv:2602.16855*, 2026.
- Tianci Xue, Weijian Qi, Tianneng Shi, Chan Hee Song, Boyu Gou, Dawn Song, Huan Sun, and Yu Su. An illusion of progress? assessing the current state of web agents. *arXiv preprint arXiv:2504.01382*, 2025.
- JingJing Xie Xuehui Wang, Zhenyu Wu et al. Mmbench-gui: Hierarchical multi-platform evaluation framework for gui agents. *arXiv preprint arXiv:2507.19478*, 2025.
- Haolong Yan, Jia Wang, Xin Huang, Yeqing Shen, Ziyang Meng, Zhimin Fan, Kaijun Tan, Jin Gao, Lieyu Shi, Mi Yang, et al. Step-gui technical report. *arXiv preprint arXiv:2512.15431*, 2025.
- Yan Yang, Dongxu Li, Yutong Dai, Yuhao Yang, Ziyang Luo, Zirui Zhao, Zhiyuan Hu, Junzhe Huang, Amrita Saha, Zeyuan Chen, Ran Xu, Liyuan Pan, Caiming Xiong, and Junnan Li. Gta1: Gui test-time scaling agent, 2025a. URL <https://arxiv.org/abs/2507.05791>.
- Yuhao Yang, Zhen Yang, Zi-Yi Dou, Anh Nguyen, Keen You, Omar Attia, Andrew Szot, Michael Feng, Ram Ramrakhya, Alexander Toshev, et al. Ultracua: A foundation model for computer use agents with hybrid action. *arXiv preprint arXiv:2510.17790*, 2025b.
- Bowen Ye, Rang Li, Qibin Yang, Yuanxin Liu, Linli Yao, Hanglong Lv, Zhihui Xie, Chenxin An, Lei Li, Lingpeng Kong, Qi Liu, Zhifang Sui, and Tong Yang. Claw-Eval: Towards trustworthy evaluation of autonomous agents. *arXiv preprint arXiv:2604.06132*, 2026. URL <https://arxiv.org/abs/2604.06132>.
- Mengqi Yuan, Zilong Zhou, Xinzhuang Xiong, Weiming Wu, Jiayang Sun, Jiamin Song, Kaiqian Cui, Bowen Wang, Haoyuan Wu, Yitong Li, Dunjie Lu, Haikong Lu, Qi Zhen, Xinyuan Wang, Jiaqi Deng, Yuhao Yang, Cheng Chen, Boyuan Zheng, Alex Su, Xiao Yu, Hao Zou, Saaket Agashe, Xing Han Lu, Manpreet Kaur, Zhengyang Qi, Vincent Sunn Chen, Frederic Sala, Dayiheng Liu, Junyang Lin, Zhou

-
- Yu, Yu Su, Siva Reddy, Xin Eric Wang, Peng Qi, Tianbao Xie, and Tao Yu. Osworld 2.0: Benchmarking computer use agents on long-horizon real-world tasks. *arXiv preprint arXiv:2606.29537*, 2026.
- Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhui Chen, and Graham Neubig. MMMU-Pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186, 2025. URL <https://arxiv.org/abs/2409.02813>.
- Z.ai. GLM-4.7: Advancing the coding capability. Z.ai Technical Blog, December 2025. URL <https://z.ai/blog/glm-4.7>. Released: 2025-12-22; accessed: 2026-07-28.
- Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. Gpt-4v (ision) is a generalist web agent, if grounded. *arXiv preprint arXiv:2401.01614*, 2024.
- Hanzhang Zhou, Xu Zhang, Panrong Tong, Jianan Zhang, Liangyu Chen, Quyu Kong, Chenglin Cai, Chen Liu, Yue Wang, Jingren Zhou, and Steven Hoi. MAI-UI Technical Report: Real-World Centric Foundation GUI Agents. *arXiv preprint arXiv:2512.22047*, 2025a. URL <https://arxiv.org/abs/2512.22047>.
- Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023. URL <https://arxiv.org/abs/2311.07911>.
- Peilin Zhou, Bruce Leon, Xiang Ying, Can Zhang, Yifan Shao, Qichen Ye, Dading Chong, Zhiling Jin, Chenxuan Xie, Meng Cao, Yuxin Gu, Sixin Hong, Jing Ren, Jian Chen, Chao Liu, and Yining Hua. BrowseComp-ZH: Benchmarking web browsing ability of large language models in Chinese. *arXiv preprint arXiv:2504.19314*, 2025b. URL <https://arxiv.org/abs/2504.19314>.
- Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. Webarena: A realistic web environment for building autonomous agents. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://arxiv.org/abs/2307.13854>.

A Appendix

A.1 Validating AutoJudge for Real-Device Evaluation

Unlike sandbox benchmarks, real-device tasks generally cannot rely on deterministic, state-based verifiers. Third-party Apps do not expose their internal states, while task outcomes may depend on changing online content, account status, or interactions across multiple Apps. Manually reviewing every trajectory is also difficult to scale and can introduce variation between annotators. We therefore validate AutoJudge against independent human annotations before using it for large-scale real-device evaluation.

Table 14: **Agreement between AutoJudge and independent human annotations.** Rows denote AutoJudge decisions and columns denote human annotations. `pass` corresponds to the combined `perfect/fair_enough` category. Agreement is computed after excluding human annotations marked as `unclear`.

AutoJudge decision	Human annotation				Class agreement
	Pass	Model failure	Environment error	Unclear	
<code>pass</code>	455	17	1	1	96.2%
<code>failed</code>	15	144	9	1	85.7%
<code>env_error</code>	1	5	19	0	76.0%
Overall	618 agreements among 666 conclusive cases				92.8%

Table 14 reports the comparison. AutoJudge agrees with the human annotations on 618 of the 666 trajectories with a conclusive human label, corresponding to an overall agreement of 92.8%. Agreement reaches 96.2% for successful trajectories. Model failures and environment errors are more difficult to distinguish, since similar observations can result from either an incorrect Agent action or a problem with the live execution environment.

We further ask expert reviewers to examine the cases in which AutoJudge and the initial human annotation disagree. Expert adjudication more often supports the AutoJudge decision, indicating that the agreement reported above is a conservative estimate of its accuracy. Based on this validation, we use AutoJudge as the primary evaluator for our real-device experiments. It provides a more accurate and scalable evaluation protocol while separately identifying model failures and errors caused by the live environment.

A.2 AutoJudge Decisions on Real-Device GUI Agent Trajectories

Figures 22–24 show representative examples of the three AutoJudge outcomes. Each example includes the original Chinese task and its English translation, selected trajectory steps with actions rendered on the screenshots, and the evidence used for the final decision. These examples illustrate how AutoJudge distinguishes successful execution and model failures from errors caused by the live execution environment.

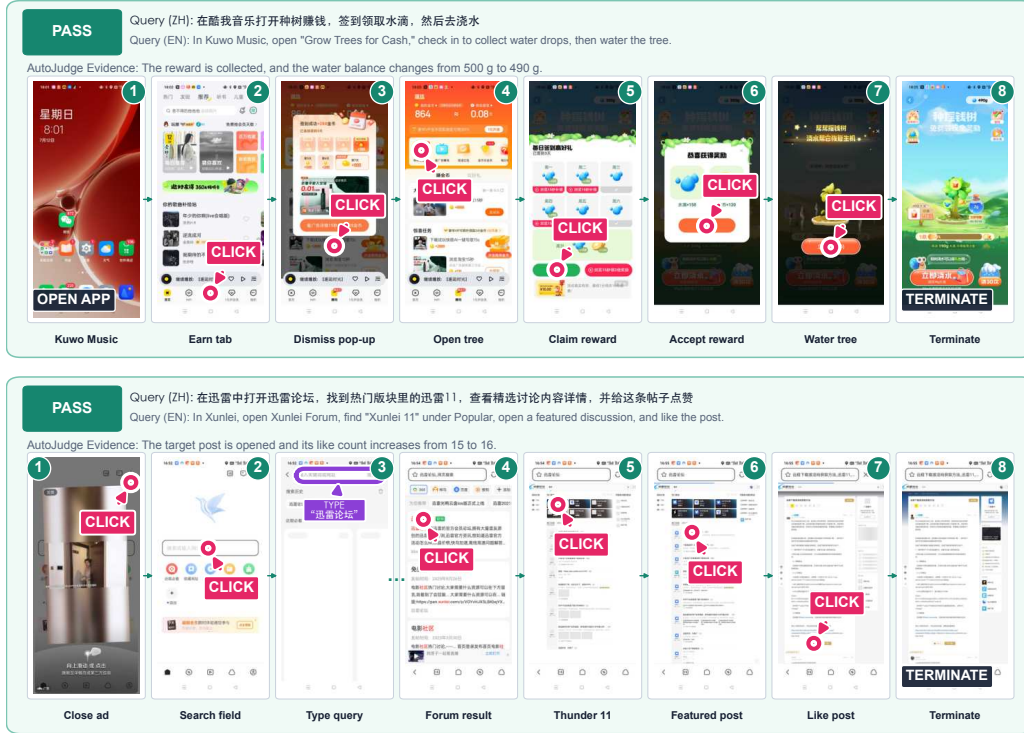


Figure 22: **Representative AutoJudge pass decisions.** The trajectory evidence shows that the Agent successfully completes the requested task.

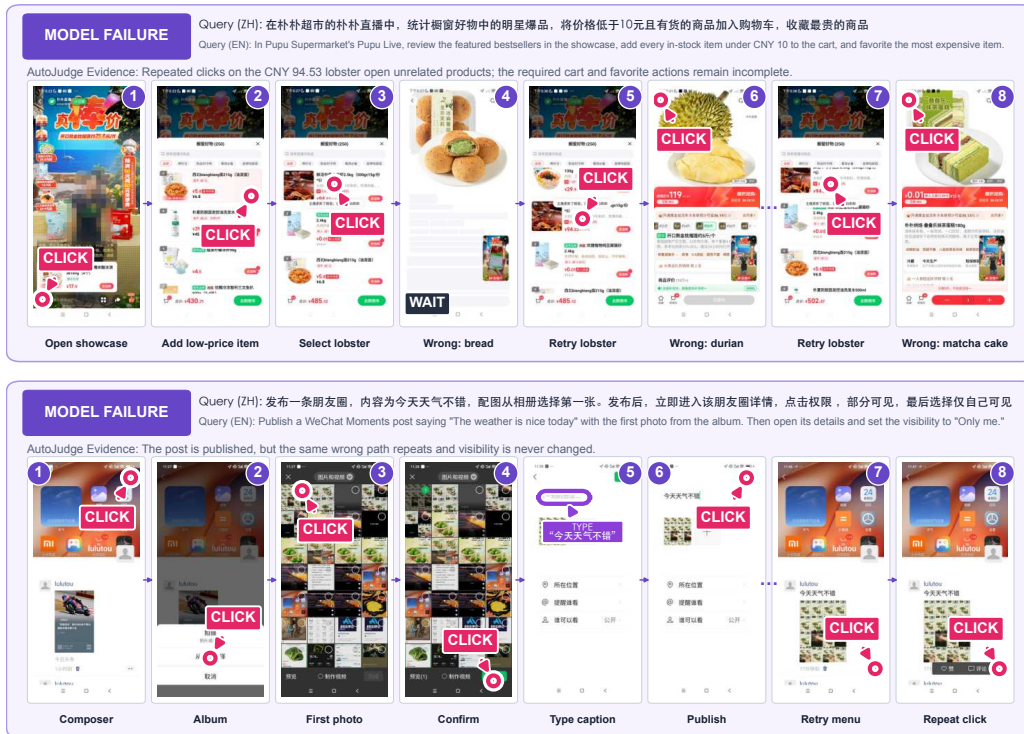


Figure 23: **Representative AutoJudge model-failure decisions.** The Agent follows an incorrect or incomplete execution path and does not satisfy the task requirements.

A.3 Additional Qualitative Examples



Figure 24: **Representative AutoJudge environment-error decisions.** The trajectories are interrupted by live-App or execution-environment issues rather than by the Agent’s actions.

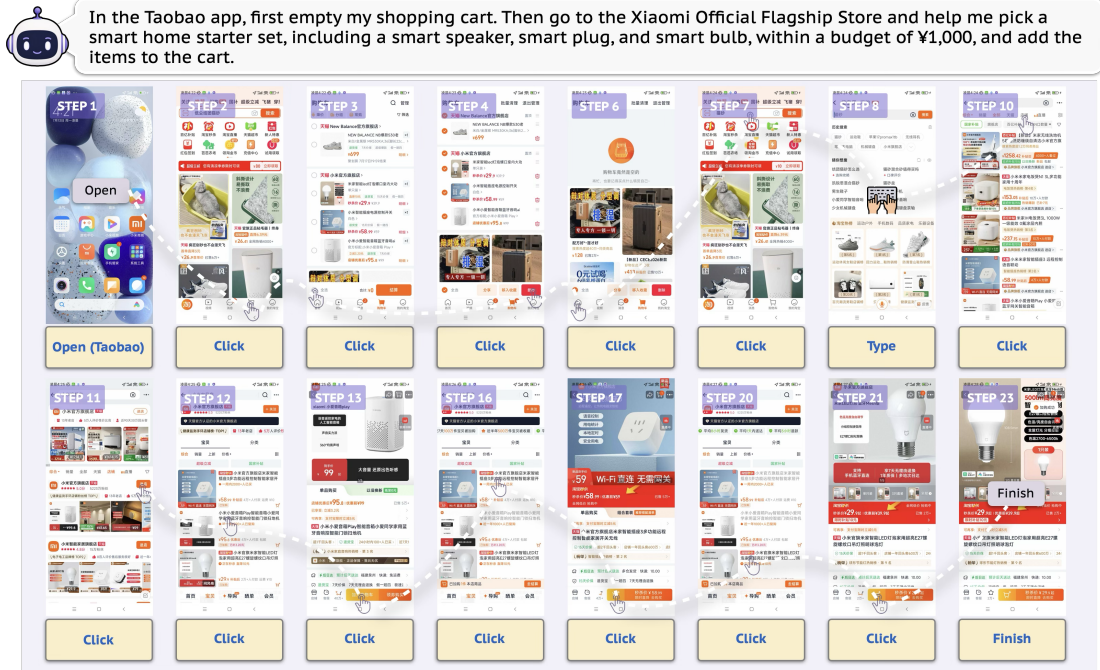


Figure 25: **Demonstration of long-horizon real-device GUI execution.** The trajectory is rendered as key frames with the executed action annotated beneath each frame. In this 23-step shopping task, the agent decomposes the instruction into sequential sub-goals: it first empties the Taobao shopping cart, then selects a smart speaker, a smart plug, and a smart bulb from the Xiaomi flagship store within a 1,000-yuan budget, maintaining the budget constraint and the item checklist throughout.



Transfer the data in the current LibreOffice Calc sheet into a LibreOffice Writer table, preserve the Calc formatting, and save the result as price.docx on the desktop.

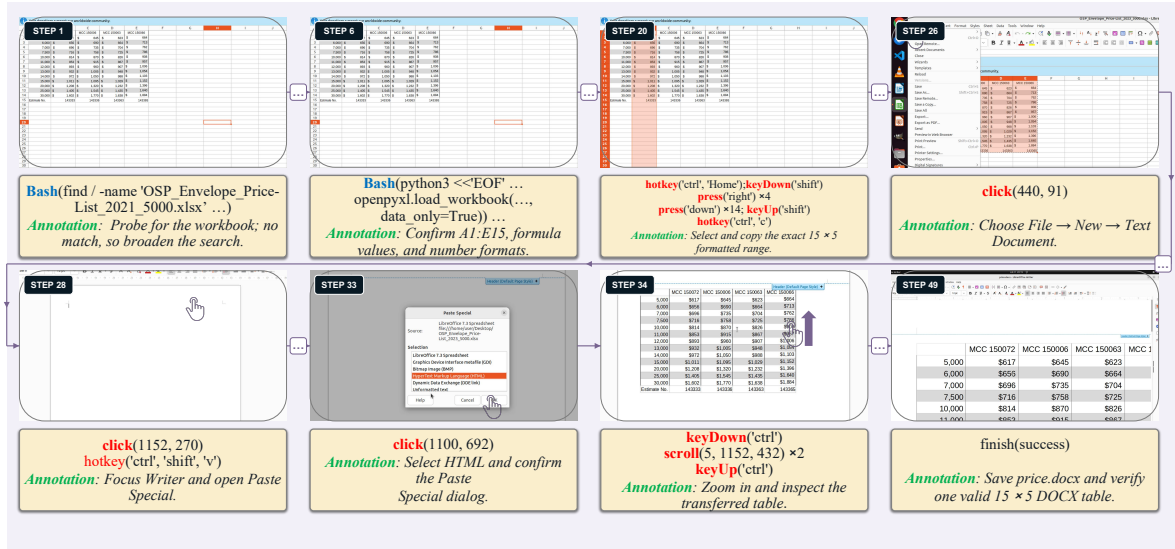


Figure 26: **Demonstration of hybrid GUI+CLI execution in computer-use tasks.** In this example, the agent uses CLI tools to inspect spreadsheet contents and interleaves GUI actions, including batched hotkey sequences and Paste Special, to transfer the formatted data into a Writer document.



First, open Douyin and find a video discussing evidence-based weight loss, then identify its main claims. Search for relevant research papers, official sources, and authoritative health information to determine whether the video contains any false or misleading statements or omits important conditions and caveats. After completing the verification, return to Douyin. If the video contains false or misleading information, write a comment that identifies the specific issue and briefly explains the supporting evidence. If the content is largely accurate, write a comment that acknowledges its main points while adding any necessary conditions or caveats

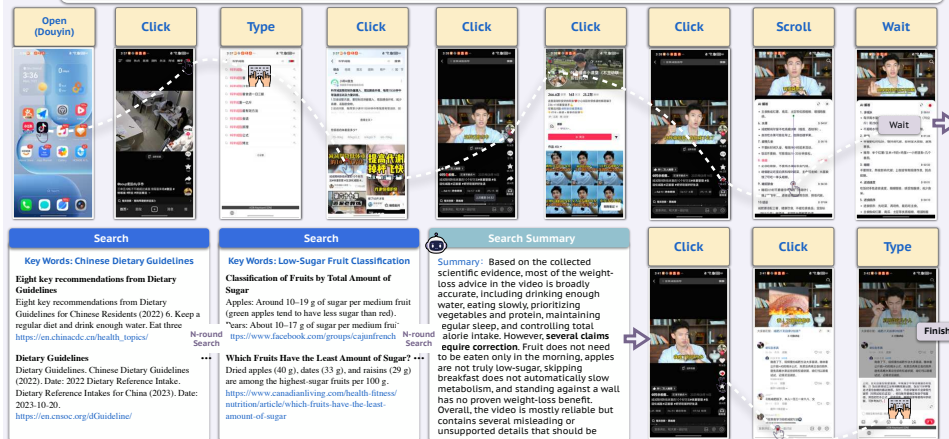


Figure 27: **Demonstration of DeepSearch invoked on demand during GUI execution.** Qwen-UI-Agent first locates a Douyin video on evidence-based weight loss and extracts its main claims through GUI actions, then calls DeepSearch mid-trajectory to verify these claims against research papers and authoritative health sources, supplying external evidence that the interface itself does not provide. The verification conclusions directly determine the subsequent GUI actions, and the agent returns to Douyin to post an evidence-grounded comment, closing an observe-research-decide-act loop.

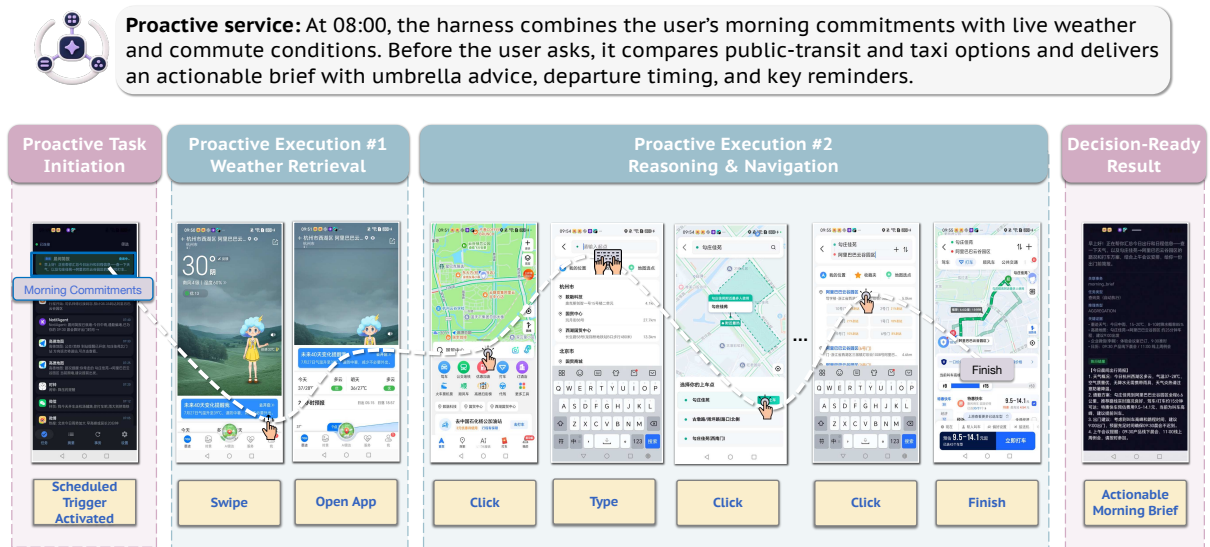


Figure 28: **Demonstration of proactive service based on mobile notifications.** The trajectory is organized into highlighted stages: proactive task initiation from a detected notification, proactive execution phases, and a decision-ready result, with the executed action annotated beneath each key frame. A scheduled morning brief: at 08:00 the harness combines the user's morning commitments with live weather and commute conditions, compares public-transit and taxi options before the user asks, and delivers an actionable brief with umbrella advice, departure timing, and key reminders.

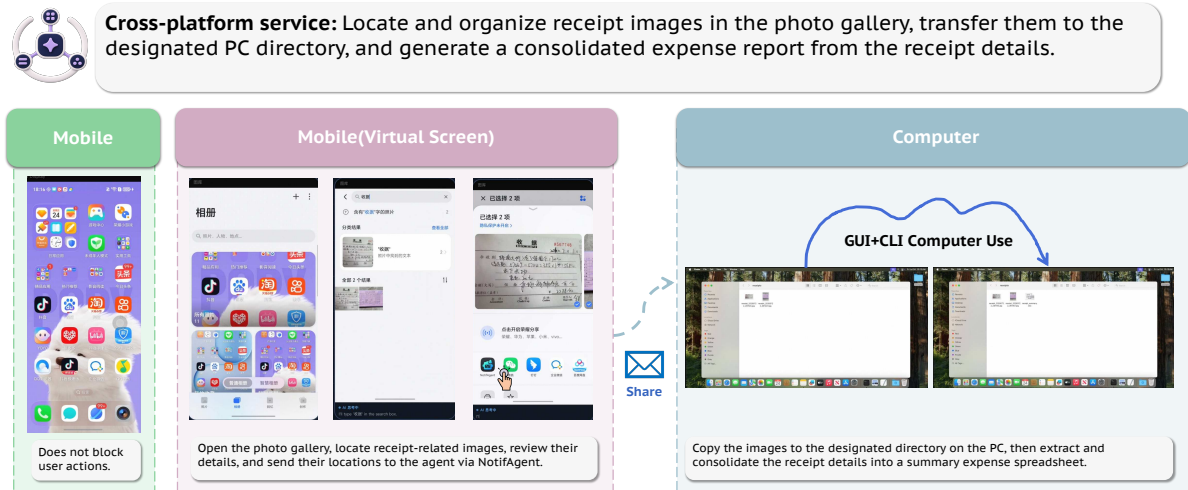


Figure 29: **Demonstrations of cross-platform task execution.** In this workflow, mobile subtasks run on virtual screens of the physical device, so execution does not block the user's own actions. Receipt organization across devices: the agent locates receipt images in the photo gallery on a mobile virtual screen, transfers them to the designated PC directory, and uses GUI+CLI computer use to consolidate the receipt details into a summary expense spreadsheet.