

# ClawGUI: One Framework to Train, Evaluate, and Deploy GUI Agents on Real Devices

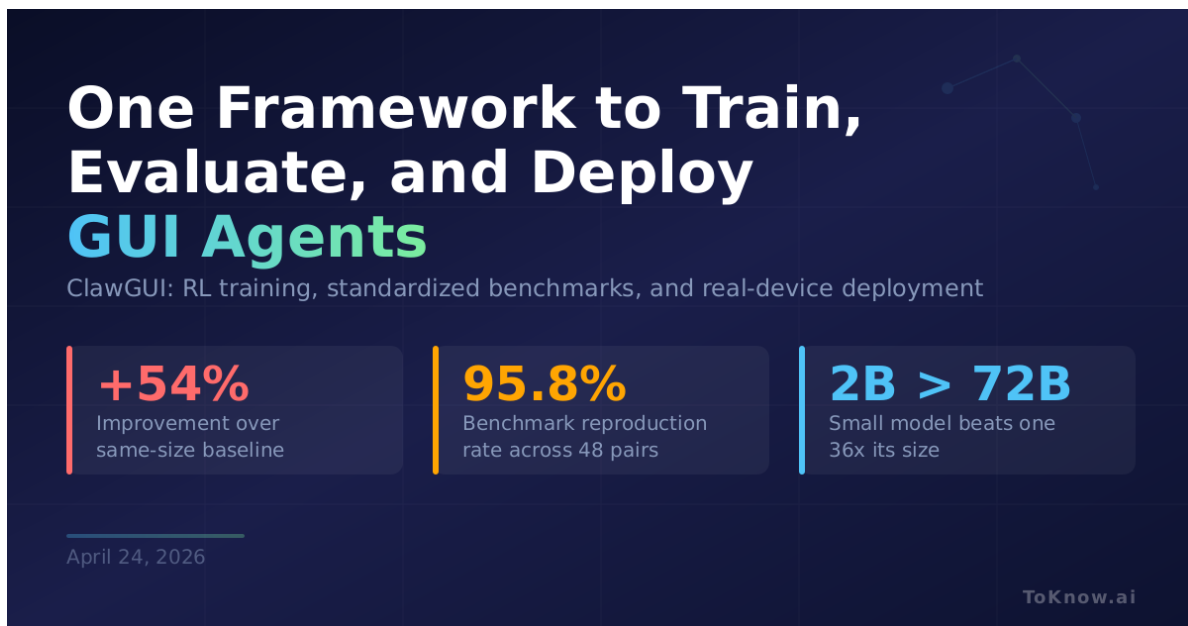
Kabui, Charles

2026-04-24

---

[Read at ToKnow.ai](#)

---



The graphic features a dark blue background with a grid pattern and a faint line graph in the top right corner. The main title is in large white and teal font. Below it, a subtitle in white text reads 'ClawGUI: RL training, standardized benchmarks, and real-device deployment'. Three key performance indicators are highlighted in separate boxes with vertical bars: '+54%' in orange for 'Improvement over same-size baseline', '95.8%' in yellow for 'Benchmark reproduction rate across 48 pairs', and '2B > 72B' in blue for 'Small model beats one 36x its size'. The date 'April 24, 2026' is at the bottom left, and the 'ToKnow.ai' logo is at the bottom right.

## One Framework to Train, Evaluate, and Deploy GUI Agents

ClawGUI: RL training, standardized benchmarks, and real-device deployment

- +54%** Improvement over same-size baseline
- 95.8%** Benchmark reproduction rate across 48 pairs
- 2B > 72B** Small model beats one 36x its size

April 24, 2026

ToKnow.ai

Zhejiang University released [ClawGUI](#), an open-source framework covering the full lifecycle of GUI agents (AI that controls apps by tapping, swiping, and typing on real screens): RL training, standardized evaluation, and deployment to real devices. ClawGUI-RL runs dozens of Docker-based Android emulators in parallel for online training and also supports training on physical phones. It pairs GiGPO with a [Process Reward Model](#) to score each interaction step

individually, so the model gets feedback on every tap and swipe rather than just a pass/fail at the end of a multi-step task. ClawGUI-Eval standardizes evaluation across 6 benchmarks and 11+ models, reproducing [95.8%](#) of officially published scores. ClawGUI-Agent deploys trained agents to Android, HarmonyOS, and iOS via 12+ chat platforms with persistent personalized memory. ClawGUI-2B, a 2B-parameter model trained entirely within this pipeline, reaches 17.1% success rate on MobileWorld GUI-Only, a 54% relative gain over the same-size MAI-UI-2B baseline and higher than UI-Venus-72B, a model 36x its size.

Standard RL for GUI tasks (GRPO) gives a single reward at the end of a 50-step phone interaction, so a wrong tap early on gets the same credit as the correct final action. GiGPO clusters steps that reach the same screen state and compares outcomes, giving the optimizer per-step signal. This alone boosted success rate from 14.5% to 17.1%. The 95.8% reproduction rate across 48 model-benchmark pairs means published GUI agent scores are finally comparable across papers, something the field badly needed.

Infrastructure matters more than model scale for GUI agents. A 2B model with dense step-level supervision and stable parallel training outperforms a 72B model without it. As training frameworks mature, the bottleneck shifts from building capable models to engineering reliable pipelines. ServiceNow’s [CUA-Suite](#) provides complementary desktop-focused data for this kind of training.

Sources:

- [ClawGUI Paper \(arXiv\)](#)
- [ClawGUI GitHub Repository](#)
- [ClawGUI Project Page](#)
- [ClawGUI-2B Model \(HuggingFace\)](#)

---

*Disclaimer: For information only. Accuracy or completeness not guaranteed. Illegal use prohibited. Not professional advice or solicitation. **Read more:** [/terms-of-service](#)*