UI-TARS: Pioneering Automated GUI Interaction with Native Agents

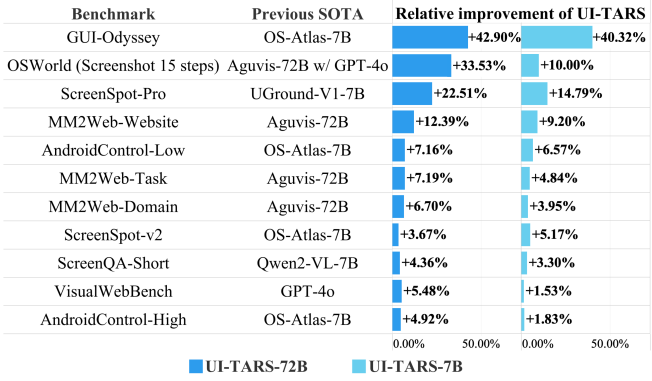
Yujia Qin[†*], Yining Ye[*◇], Junjie Fang[*], Haoming Wang[*], Shihao Liang[*], Shizuo Tian[◇], Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, Wanjun Zhong, Kuanye Li, Jiale Yang, Yu Miao, Woyu Lin, Longxiang Liu, Xu Jiang, Qianli Ma, Jingyu Li, Xiaojun Xiao, Kai Cai, Chuang Li, Yaowei Zheng, Chaolin Jin, Chen Li, Xiao Zhou, Minchao Wang, Haoli Chen, Zhaojian Li, Haihua Yang, Haifeng Liu, Feng Lin, Tao Peng, Xin Liu, Guang Shi[†]

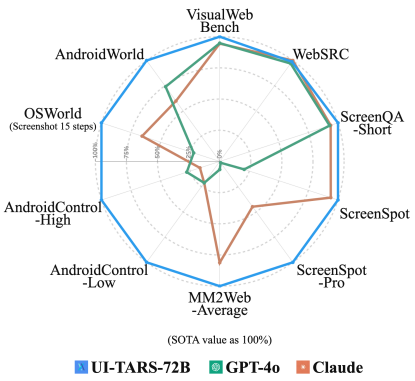
ByteDance Seed, [◇]Tsinghua University

{yujia.qin, shiguang.sg}@bytedance.com

Abstract

This paper introduces UI-TARS, a native GUI agent model that solely perceives the screenshots as input and performs human-like interactions (e.g., keyboard and mouse operations). Unlike prevailing agent frameworks that depend on heavily wrapped commercial models (e.g., GPT-4o) with expert-crafted prompts and workflows, UI-TARS is an end-to-end **model** that outperforms these sophisticated **frameworks**. Experiments demonstrate its superior performance: UI-TARS achieves SOTA performance in 10+ GUI agent benchmarks evaluating perception, grounding, and GUI task execution (see below). Notably, in the OSWorld benchmark, UI-TARS scores 22.7 with 15 steps, outperforming Claude's 14.9; in AndroidWorld, UI-TARS achieves 46.6, surpassing GPT-4o's 34.5. UI-TARS incorporates several key innovations: (1) **Enhanced Perception**: leveraging a large-scale dataset of GUI screenshots for context-aware understanding of UI elements and precise captioning; (2) **Unified Action Modeling**, which standardizes actions into a unified space to improve multi-step task execution across platforms, achieving precise grounding and interaction through large-scale action traces; (3) **System-2 Reasoning**, which incorporates deliberate reasoning into multi-step decision making, involving multiple reasoning patterns such as task decomposition, reflection thinking, milestone recognition, etc. (4) **Iterative Training with Reflective Online Traces**, which addresses the data bottleneck by automatically collecting, filtering, and reflectively refining new interaction traces on hundreds of virtual machines. Through iterative training and reflection tuning, UI-TARS continuously learns from its mistakes and adapts to unforeseen situations with minimal human intervention. We also analyze the evolution path and core capability of GUI agents to guide the further development of this domain. UI-TARS is open sourced at https://github.com/bytedance/UI-TARS.





[*]Indicates equal contribution. [†] Corresponding author. [◇] The work is done when interning at ByteDance.