GUI Agents with Foundation Models: Data Resource, Framework and Application

Shuai Wang[1] , **Kaiwen Zhou**[1] , **Rui Shao**[2] , **Gongwei Chen**[2] , **Yuqi Zhou**[3]

[1]Huawei Noah's Ark Lab [2]Harbin Institute of Technology, Shenzhen [3]Renmin University of China

1 Abstract

The rapid advancement of foundation models like large vision language models (VLMs) has paved the way for intelligent agents capable of autonomously interacting with Graphical User Interfaces (GUIs). This tutorial provides a comprehensive overview of the latest innovations in GUI agents and influential work across data resource, framework, and application.

2 Tutorial description

This tutorial aims to provide a structured overview of the latest work in the field of GUI agents. As depicted in Figure 1, we deconstructs GUI agent ecosystems into three critical pillars: multimodal data resources, algorithmic frameworks, and applications: Data resources, such as user instructions, user interface (UI) screenshots, and behavior traces, form the cornerstone for the architecture design of GUI agent [Rawles *et al.*, 2023; Lu *et al.*, 2024]; Frameworks orchestrate the power of foundation models, knowledge bases and tools to enable intelligent and reliable decision-making [Li *et al.*, 2024a; Wang *et al.*, 2024a; Zhu *et al.*, 2024]; Applications represent the concrete setups for domain-optimized implementation [Lai *et al.*, 2024; Liu *et al.*, 2024]. The current state of these aspects reflects the maturity of the field and highlights future research priorities.

To this end, we organize this tutorial to be a captivating review and lecture around three key areas: **Data Resources**, **Frameworks**, and **Applications**.

3 Tutorial Content

3.1 Target Audience

The tutorial is developed for researchers and engineers who are interested in GUI Agents.

This tutorial balances the introductory and advanced material, i.e., 50 % for beginners and 50 % for intermediate and advanced researchers.

In addition, as this tutorial will be an organic combination of (1) the application of GUI agents, as well as (2) the algorithmic advancement of GUI agents in their framework and grounding models, we believe that both the business practitioner as well as researchers focusing on algorithmic aspects will benefit from our contents.

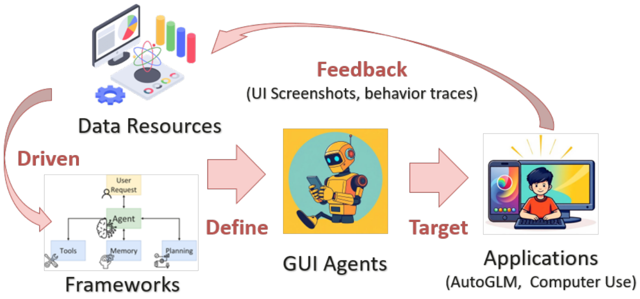

Figure 1: The foundational aspects and goals of GUI agents.

3.2 Prerequisites

As to prerequisites, basic background of Large Language Models and Automatic Agents would be sufficient. Since we will introduce the basic concepts of both topics, a background in them will be helpful but is not a strict requirement.

3.3 Agenda

This will be **a 1/2 day tutorial** to cover the state-of-the-art research for GUI Agents with theories and applications.

- **Part I: Introduction**
  Duration: 15 minutes, *Presenter:* Shuai Wang
  We start by motivating the formulation of GUI agents with foundation models and their existing challenges.
  - Motivation
  - Formulation of GUI Agents [Wang *et al.*, 2024c]
  - Challenges
- **Part II: GUI Agents Data Resource**
  Duration: 30 minutes, *Presenter:* Kaiwen Zhou and Yuqi Zhou
  This is the core part, we focus on recent datasets and benchmarks to train and evaluate the capabilities of VLM- based GUI agents.
  - Datasets [Rawles *et al.*, 2023; Lin *et al.*, 2024]
  - Environments [Rawles *et al.*, 2024]
  - Evaluation Benchmarks [Chen *et al.*, 2024c]
- **Part III: GUI Agents Construction**
  Duration: 60 minutes, Presenter: Rui Shao and Gongwei Chen

This is the core part, we delve into the innovative framework of GUI Agents, especially focusing on recent advancements in the grounding models and multi-agent frameworks.

  - Grounding Models [Yang *et al.*, 2024; Gou *et al.*, 2024; Lin *et al.*, 2024]
  - Multi-Agent Frameworks [Wang *et al.*, 2024b; Zhu *et al.*, 2024; Zhou *et al.*, 2025]

- **Part IV: GUI Agents Applications**
  Duration: 30 minutes, Presenter: Shuai Wang
  In this part, we discuss the commercial applications widely used in industrial settings.
  - Apple Intelligence
  - OpenAI Operator
  - AutoGLM [Liu *et al.*, 2024]
- **Part V: Open Questions and Future Trends**
  Duration: 30 minutes, Presenter: Rui Shao
  In this part, we summarize and discuss several key research questions that require urgent attentions in this field.
  - Personalized GUI Agents
  - Security of GUI Agents
  - Inference Efficiency

4 Short Bio

In this section, we briefly introduce the 5 presenters of this tutorial.

Shuai Wang is a senior researcher at Huawei Noah's Ark Lab. He received his Ph.D. from Peking University in 2021. Currently, his research focuses on the basic research and applications of GUI agents [Zhou *et al.*, 2025; Chen *et al.*, 2024c; Wang *et al.*, 2024c]. Dr. Wang's work has contributed to the cutting-edge industrial applications. He has been published papers in top machine learning conference(e.g., WWW, ICLR). Dr. Wang served as a teaching assistant for AI-related courses at Peking University. While at Huawei, he delivered numerous reports and internal tutorial sessions on GUI Agent topics.

Kaiwen Zhou is a senior researcher at Huawei Noah's Ark Lab. He received his Ph.D. and M.Phil. from the Chinese University of Hong Kong, and B.S. from Fudan University. He is currently interested in research topics related to agent [Chen *et al.*, 2024c; Chen *et al.*, 2024b], optimization for machine learning and out-of-distribution generalization of machine learning solutions. Dr. Zhou has published more than 10 papers in top conferences and journals in arificial intelligence such as NeurIPS, ICML, ICLR, AISTATS, IJCAI, AAAI, etc. He also served as a program committee member at NeurIPS, ICML, ICLR, AISTATS.

Rui Shao is currently a professor with Harbin Institute of Technology (Shenzhen). He received his B.Eng. and Ph.D. degree from University of Electronic Science and Technology of China in 2015 and Hong Kong Baptist University in 2021, respectively. His research interests lie primarily in multimodal learning and computer vision. He has published more than 30 first/corresponding author papers in top journals and conferences, such as TPAMI, IJCV, CVPR, ECCV. He serves as a reviewer for many top journals and conferences, including TPAMI, IJCV, CVPR, ICCV, ECCV. He serves as an Area Chair of ACM MM, BMVC, and Panel Co-Chairs of ICMR. His current research interests include multimodal large language models (MLLMs) [Chen *et al.*, 2024a; Shen *et al.*, 2024] and agents [Chen *et al.*, 2024c; Li *et al.*, 2024b; Li *et al.*, 2025]. He has taught several computer science lessons, such as Computer Network.

Gongwei Chen is currently a postdoc at Harbin Institute of Technology (Shenzhen). He received his Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences (ICT, CAS), and his B.Eng. degree from the University of Science and Technology Beijing (USTB). His current research interests include multimodal large language models (MLLMs) [Chen *et al.*, 2024a; Shen *et al.*, 2024] and MLLM-based agents [Chen *et al.*, 2024c; Li *et al.*, 2024b; Li *et al.*, 2025]. He has published more than 10 papers in top conferences and journals, including CVPR, NeurIPS, ICLR, ACM MM, TIP, etc.

Yuqi Zhou is currently a second-year Ph.D. candidate at the Gaoling School of Artificial Intelligence, Renmin University of China. He received his Bachelor's degree in Computer Science and Technology from Harbin Institute of Technology (Shenzhen). His research focuses on GUI Agents and their applications in mobile agent construction. He has published papers in top-tier conferences, including KDD, ACL, and CIKM, contributing to advancements in artificial intelligence and human-computer interaction.

References

[Chen *et al.*, 2024a] Gongwei Chen, Leyang Shen, Rui Shao, Xiang Deng, and Liqiang Nie. Lion: Empowering multimodal large language model with dual-level visual knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26540–26550, 2024.

[Chen *et al.*, 2024b] Haoyu Chen, Wenbo Li, Jinjin Gu, Jingjing Ren, Sixiang Chen, Tian Ye, Renjing Pei, Kaiwen Zhou, Fenglong Song, and Lei Zhu. Restoreagent: Autonomous image restoration agent via multimodal large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 110643–110666, 2024.

[Chen *et al.*, 2024c] Jingxuan Chen, Derek Yuen, Bin Xie, et al. Spa-bench: A comprehensive benchmark for smartphone agent evaluation. In *NeurIPS 2024 Workshop on Open-World Agents*, 2024.

[Gou *et al.*, 2024] Boyu Gou, Ruohan Wang, Boyuan Zheng, Yanan Xie, Cheng Chang, Yiheng Shu, Huan Sun, and Yu Su. Navigating the digital world as humans do: Universal visual grounding for gui agents. *arXiv preprint arXiv:2410.05243*, 2024.

[Lai *et al.*, 2024] Hanyu Lai, Xiao Liu, Iat Long Iong, et al. Autowebglm: A large language model-based web navigating agent. In *SIGKDD*, pages 5295–5306, 2024.

[Li *et al.*, 2024a] Yanda Li, Chi Zhang, Wanqi Yang, et al. Appagent v2: Advanced agent for flexible mobile interactions. *arXiv preprint arXiv:2408.11824*, 2024.

[Li *et al.*, 2024b] Zaijing Li, Yuquan Xie, Rui Shao, Gongwei Chen, Dongmei Jiang, and Liqiang Nie. Optimus-1: Hybrid multimodal memory empowered agents excel in long-horizon tasks. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[Li *et al.*, 2025] Zaijing Li, Yuquan Xie, Rui Shao, Gongwei Chen, Dongmei Jiang, and Liqiang Nie. Optimus-2: Multimodal minecraft agent with goal-observation-action conditioned policy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[Lin *et al.*, 2024] Kevin Qinghong Lin, Linjie Li, Difei Gao, Zhengyuan Yang, Shiwei Wu, Zechen Bai, Weixian Lei, Lijuan Wang, and Mike Zheng Shou. Showui: One vision-language-action model for gui visual agent. *arXiv preprint arXiv:2411.17465*, 2024.

[Liu *et al.*, 2024] Xiao Liu, Bo Qin, Dongzhu Liang, et al. Autoglm: Autonomous foundation agents for guis. *arXiv preprint arXiv:2411.00820*, 2024.

[Lu *et al.*, 2024] Quanfeng Lu, Wenqi Shao, Zitao Liu, et al. Gui odyssey: A comprehensive dataset for cross-app gui navigation on mobile devices. *arXiv preprint arXiv:2406.08451*, 2024.

[Rawles *et al.*, 2023] Christopher Rawles, Alice Li, Daniel Rodriguez, Oriana Riva, and Timothy P Lillicrap. Androidinthewild: A large-scale dataset for android device control. In *NIPS Datasets and Benchmarks Track*, 2023.

[Rawles *et al.*, 2024] Christopher Rawles, Sarah Clinckemaillie, Yifan Chang, et al. Androidworld: A dynamic benchmarking environment for autonomous agents. *arXiv preprint arXiv:2405.14573*, 2024.

[Shen *et al.*, 2024] Leyang Shen, Gongwei Chen, Rui Shao, Weili Guan, and Liqiang Nie. Mome: Mixture of multimodal experts for generalist multimodal large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[Wang *et al.*, 2024a] Junyang Wang, Haiyang Xu, Haitao Jia, et al. Mobile-agent-v2: Mobile device operation assistant with effective navigation via multi-agent collaboration. In *NIPS*, 2024.

[Wang *et al.*, 2024b] Junyang Wang, Haiyang Xu, Haitao Jia, Xi Zhang, Ming Yan, Weizhou Shen, Ji Zhang, Fei Huang, and Jitao Sang. Mobile-agent-v2: Mobile device operation assistant with effective navigation via multi-agent collaboration. *arXiv preprint arXiv:2406.01014*, 2024.

[Wang *et al.*, 2024c] Shuai Wang, Weiwen Liu, Jingxuan Chen, Yuqi Zhou, Weinan Gan, Xingshan Zeng, Yuhan Che, Shuai Yu, Xinlong Hao, Kun Shao, et al. Gui agents with foundation models: A comprehensive survey. *arXiv preprint arXiv:2411.04890*, 2024.

[Yang *et al.*, 2024] Yuhao Yang, Yue Wang, Dongxu Li, et al. Aria-ui: Visual grounding for gui instructions. *arXiv preprint arXiv:2412.16256*, 2024.

[Zhou *et al.*, 2025] Yuqi Zhou, Shuai Wang, Sunhao Dai, Qinglin Jia, Zhaocheng Du, Zhenhua Dong, and Jun Xu. Chop: Mobile operating assistant with constrained high-frequency optimized subtask planning. *arXiv preprint arXiv:2503.03743*, 2025.

[Zhu *et al.*, 2024] Zichen Zhu, Hao Tang, Yansi Li, et al. Moba: A two-level agent system for efficient mobile task automation. *arXiv preprint arXiv:2410.13757*, 2024.