Awesome GUI Agents: Papers, Models, and Datasets

Benchmark Results

OSWorld

Rank	Model	SR (%)	Steps	Base Model	Input	Platforms	Availability	Paper/Source	Date
1	UI-TARS-72B-DPO	24.6	50	Qwen-2-VL-72B	Screenshots	Web, Desktop, Mobile	Model open sourced, dataset and training code proprietary	UI-TARS	2025/01/21
2	UI-TARS-72B-DPO	22.7	15	Qwen-2-VL-72B	Screenshots	Web, Desktop, Mobile	Model open sourced, dataset and training code proprietary	UI-TARS	2025/01/21
3	Claude Computer-Use	22.0	50	Claude-3.5	Screenshots	Desktop	Proprietary	Anthropic	-
4	UI-TARS-7B-DPO + ZeroGUI	20.2	15	UI-TARS-7B-DPO	Screenshots	Desktop, Mobile	Training code samples open sourced	ZeroGUI	2025/05/29
5	OpenAI Operator	19.7	-	-	Screenshots	Desktop	Proprietary	OpenAI	-
6	UI-TARS-72B-SFT	18.8	15	Qwen-2-VL-72B	Screenshots	Web, Desktop, Mobile	Model open sourced, dataset and training code proprietary	UI-TARS	2025/01/21
7	UI-TARS-7B-DPO	18.1	15	Qwen-2-VL-7B	Screenshots	Web, Desktop, Mobile	Model open sourced, dataset and training code proprietary	UI-TARS	2025/01/21
8	UI-TARS-7B-SFT	17.7	15	Qwen-2-VL-7B	Screenshots	Web, Desktop, Mobile	Model open sourced, dataset and training code proprietary	UI-TARS	2025/01/21
9	Claude Computer-Use	14.9	15	Claude-3.5	Screenshots	Desktop	Proprietary	Anthropic	-
9	GLM-4.1V-9B-Thinking	14.9	-	GLM-4.1V-9B	Screenshots	Mobile, Desktop	Model: MIT License, Code: Apache License 2.0	GLM-4.1V-Thinking	2025/07/02
11	AGUVIS-72B	10.26	-	Qwen2-VL-72B	Screenshots	Web, Desktop, Mobile	Model, dataset, and training scripts open sourced	AGUVIS	2025/05/05
12	Qwen2.5-VL-72B	8.8	-	Qwen2.5-VL-72B	Screenshots	General	Qwen License	Base Model	-
13	Kimi-VL A3B-Thinking	8.2	-	-	Screenshots	-	-	-	-
14	CogAgent-9B	8.1	-	CogAgent-9B	Screenshots	General	-	-	-
15	Gemini-Pro-1.5	5.4	-	Gemini-Pro-1.5	Screenshots	General	Proprietary	Google	-
16	GPT-4o (2024-11-20)	5.0	-	GPT-4o	Screenshots+a11y tree	General	Proprietary	OpenAI	-
17	Aguvis-7B + ZeroGUI	4.9	15	Aguvis-7B	Screenshots	Desktop, Mobile	Training code samples open sourced	ZeroGUI	2025/05/29
18	Aguvis-7B	3.0	15	Qwen2-VL-7B	Screenshots	Web, Desktop, Mobile	Model, dataset, and training scripts open sourced	AGUVIS	2025/05/05
19	Qwen2.5-VL-7B	1.9	-	Qwen2.5-VL-7B	SoM	General	Qwen License	Base Model	-
19	MiMo-VL 7B-RL	1.9	-	-	Screenshots	-	-	-	-
21	InternVL3-9B	1.4	-	InternVL3-9B	Screenshots	General	Open sourced	Base Model	-

AndroidWorld

Rank	Model	SR (%)	Steps	Base Model	Input	Platforms	Availability	Paper/Source	Date
1	JT-GUIAgent-V1	60.0	-	-	-	-	-	-	-
2	Claude Computer-Use	55.0	50	Claude-3.5	General	Desktop	Proprietary	Anthropic	-
3	Agent S2	54.3	-	UI-TARS-72B-DPO	Screenshots	Desktop, Android	Inference code open sourced	Agent S2	2025/04/01
4	UI-TARS-7B-DPO + ZeroGUI	47.5	-	UI-TARS-7B-DPO	Screenshots	Desktop, Mobile	Training code samples open sourced	ZeroGUI	2025/05/29
5	GUI-Explorer (w/ SoM)	47.4	-	Qwen2-VL, Qwen2.5-VL	SoM	Mobile	Inference code open sourced	GUI-explorer	2025/05/22
6	UI-TARS-72B-SFT	46.6	15	Qwen-2-VL-72B	Screenshots	Web, Desktop, Mobile	Model open sourced, dataset and training code proprietary	UI-TARS	2025/01/21
7	UI-TARS-7B-DPO	45.7	-	Qwen-2-VL-7B	Screenshots	Web, Desktop, Mobile	Model open sourced, dataset and training code proprietary	UI-TARS	2025/01/21
8	Aria-UI	44.8	-	-	-	-	-	-	-
9	GLM-4.1V-9B-Thinking	41.7	-	GLM-4.1V-9B	Screenshots	Mobile, Desktop	Model: MIT License, Code: Apache License 2.0	GLM-4.1V-Thinking	2025/07/02
10	Aguvis-7B	37.1	-	Qwen2-VL-7B	Screenshots	Web, Desktop, Mobile	Model, dataset, and training scripts open sourced	AGUVIS	2025/05/05
11	Qwen2.5-VL-72B	35.0	-	Qwen2.5-VL-72B	Screenshots	General	Open sourced	Base Model	-
12	GPT-4o (2024-11-20)	34.5	-	GPT-4o	Screenshots+a11y tree	General	Proprietary	OpenAI	-
13	UI-TARS-7B-SFT	33.0	15	Qwen-2-VL-7B	Screenshots	Web, Desktop, Mobile	Model open sourced, dataset and training code proprietary	UI-TARS	2025/01/21
14	UGround	32.8	-	-	-	-	-	-	-
15	Claude Computer-Use	27.9	15	Claude-3.5	Screenshots	Desktop	Proprietary	Anthropic	-
16	Qwen2.5-VL-7B	27.6	-	Qwen2.5-VL-7B	Screenshots	General	Open sourced	Base Model	-
17	Aguvis-72B	26.1	-	Qwen2-VL-72B	Screenshots	Web, Desktop, Mobile	Model, dataset, and training scripts open sourced	AGUVIS	2025/05/05
18	Gemini-Pro-1.5	22.8	-	Gemini-Pro-1.5	Screenshots	General	Proprietary	Google	-
19	MiMo-VL 7B-RL	10.8	-	-	Screenshots	-	-	-	-
20	CogAgent-9B-20241220	8.1	-	CogAgent-9B	Screenshots	General	-	-	-
21	InternVL3-9B	1.9	-	InternVL3-9B	Screenshots	General	Open sourced	Base Model	-

GUI Agent Dataset Sources and Pipelines

Model/Framework	Dataset Construction Pipeline	Data Source	Construction Method	Code Availability	Dataset Size/Type
MONDAY	✅ **Yes**	Video tutorials	OCR-based scene detection, UI element detection, multi-step action recognition	Open sourced	Video-to-dataset conversion
TongUI	✅ **Yes**	Online tutorial videos & articles	Web crawling + trajectory data extraction	crawler code + processing open sourced	GUI-Net-1M (1M trajectory samples)
ZeroGUI	✅ **Yes**	Online learning/interaction	VLM-based automatic task generation (GPT-4o) + automatic reward estimation + RL	Training code samples open	Continuous online generation
GUI-Robust	✅ **Yes**	Real-world GUI anomalies	Semi-automated via RPA tools + YOLOv8 + Qwen2.5-VL	Dataset open sourced	Robustness testing dataset
OS-ATLAS	❌ No	Pre-existing dataset	Uses existing 13M GUI elements grounding dataset	Inference code only	13M GUI elements
UI-TARS	❌ No	Proprietary dataset	Uses proprietary ~50B tokens dataset	Model open, dataset proprietary	~50B tokens
Agent S2	❌ No	Existing models/tools	Uses UI-TARS-72B-DPO, Claude models, OCR	Inference code only	N/A
LearnAct	❌ No	Pre-built benchmark	Uses LearnGUI benchmark	Inference code + dataset open	2,252 offline + 101 online tasks
AGUVIS	❌ No	Pre-existing dataset	Uses large-scale multimodal dataset	Model + dataset + training scripts	Large-scale multimodal annotations
GUI-explorer	❌ No	Autonomous exploration	Training-free exploration + knowledge mining	Inference code only	Knowledge vector store
AgentCPM-GUI	❌ No	Pre-existing dataset	Uses CAGUI benchmark	SFT + RFT code, dataset proprietary	55K trajectories, 470K steps
GLM-4.1V-Thinking	❌ No	Closed	Uses existing GUI benchmarks	Model opened	N/A

Paper List

1. OS-ATLAS

- Title: OS-ATLAS: A Foundation Action Model for Generalist GUI Agents
- Date: 2024/10/30
- Input: Screenshots
- Base Models: InternVL2-4B → OS-Atlas-Base-4B; Qwen2-VL-7B-Instruct → OS-Atlas-Base-7B
- Dataset: 13 million GUI elements grounding dataset
- Method: GUI grounding pretraining + action fine-tuning
- Platforms: Windows, Linux, macOS, Android, and Web
- Availability: Inference code open sourced

2. UI-TARS

- Title: UI-TARS: Pioneering Automated GUI Interaction with Native Agents
- Date: 2025/01/21

- Base Models: Qwen-2-VL-7B → UI-TARS-7B; Qwen-2-VL-72B → UI-TARS-72B
- Dataset: ~50 billion tokens (proprietary)
- Platforms: Web, Desktop, Mobile
- Method:
  - Three-phase training: Continual Pre-training Phase, Annealing Phase, and DPO Phase
  - Enhanced perception + unified action modeling + system-2 reasoning + iterative training using reflective online traces
- Performance Reported:
  - OSWorld:
    - GPT-4o: 5.0
    - Gemini-Pro-1.5: 5.4
    - Claude Computer-Use (15 steps): 14.9
    - Claude Computer-Use (50 steps): 22.0
    - Aguvis-72B: 10.3
    - UI-TARS-7B-SFT (15 steps): 17.7
    - UI-TARS-7B-DPO (15 steps): 18.1
    - UI-TARS-72B-SFT (15 steps): 18.8
    - UI-TARS-72B-DPO (15 steps): 22.7
    - UI-TARS-72B-DPO (50 steps): 24.6
  - AndroidWorld:
    - GPT-4o: 34.5 (SoM)
    - Gemini-Pro-1.5: 22.8 (SoM)
    - CogAgent-9B-20241220: 8.1
    - Claude Computer-Use (15 steps): 27.9
    - Claude Computer-Use (50 steps): 55.0
    - Aguvis-72B: 26.1
    - UI-TARS-7B-SFT (15 steps): 33.0
    - UI-TARS-72B-SFT (15 steps): 46.6
    - UGround: 32.8
    - Aria-UI: 44.8
    - Aguvis-7B: 37.1
- Availability: Model open sourced, dataset and training code proprietary

3. Agent S2

- Title: Agent S2: A Compositional Generalist-Specialist Framework for Computer Use Agents
- Date: 2025/04/01
- Input: Screenshots
- Method:
  - Mixture of Grounding for resolving the grounding bottleneck
  - Proactive Hierarchical Planning for dynamic replanning
- Models Used: UI-TARS-72B-DPO, Claude-3.7, Claude-3.5, Tesseract OCR, Universal Network Objects (UNO)
- Performance Reported:
  - AndroidWorld:

        - Agent S2: 54.3%
        - UI-TARS-72B-SFT: 46.6%
- Platforms: Desktop, Android
- Availability: Inference code open sourced

4. LearnAct

- Title: LearnAct: Few-Shot Mobile GUI Agent with a Unified Demonstration Benchmark
- Date: 2025/04/18
- Input: Screenshots
- Dataset: LearnGUI, 2,252 offline tasks and 101 online tasks
- Method:
    - LearnAct multi-agent framework, automatically extracts knowledge from demonstrations:
        - DemoParser for knowledge extraction
        - KnowSeeker for retrieval of relevant knowledge
        - ActExecutor for demonstration-enhanced task execution
- Performance Reported: Improves Performance Reported of Gemini, UI-TARS-7B-SFT, and Qwen2-VL-7B on LearnGUI datasets
- Platforms: Mobile
- Availability: Inference code and dataset open sourced

5. AGUVIS

- Title: AGUVIS: Unified Pure Vision Agents for Autonomous GUI Interaction
- Date: 2025/05/05
- Input: Screenshots
- Dataset: Large-scale dataset with multimodal grounding and reasoning annotations
- Method: Two-stage training pipeline separating GUI grounding from planning and reasoning
- Base Model: Qwen2-VL
- Performance Reported:
    - OSWorld:
        - AGUVIS-72B: 10.26
        - Claude Computer-Use: 14.9
        - OpenAI Operator: 19.7
    - AndroidWorld:
        - AGUVIS-72B: 26.1
        - GPT-4o Planner + AGUVIS-7B Grounder: 37.1
- Platforms: Web, Desktop, Mobile
- Availability: Model, dataset, and training scripts open sourced

6. MONDAY

- Title: Scalable Video-to-Dataset Generation for Cross-Platform Mobile Agents
- Date: 2025/05/19
- Method:
    - A framework that converts video to dataset without manual annotation

    - OCR-based scene detection, UI element detection, and multi-step action recognition
- Platforms: Android, iOS
- Availability: MONDAY dataset and data processing code open sourced

7. TongUI

- Title: TongUI: Building Generalized GUI Agents by Learning from Multimodal Web Tutorials
- Date: 2025/05/21
- Input: Screenshots
- Method:
    - Crawls online tutorial videos and articles into GUI agent trajectory data
    - Constructs GUI-Net-1M dataset with 1 million trajectory samples
    - Fine-tuned Qwen2.5-VL-3B/7B models on GUI-Net-1M dataset
- Models Used: Qwen2.5-VL-3B / 7B
- Performance Reported:
    - AITW: 73.3 average
- Platforms: Mobile, Desktop
- Availability: Datasets, models, training scripts, crawler code, and intermediate data all open sourced

8. GUI-explorer

- Title: GUI-explorer: Autonomous Exploration and Mining of Transition-aware Knowledge for GUI Agent
- Date: 2025/05/22
- Input: SoM
- Method:
    - Training-free agent with:
        - Autonomous exploration of function-aware trajectories
        - Unsupervised mining of transition-aware knowledge
        - Dynamic guidance via a knowledge vector store
- Performance Reported: Improves Qwen2-VL and Qwen2.5-VL
    - AndroidWorld: 47.4%
    - SPA-Bench: 53.7%
- Platforms: Mobile
- Availability: Inference code open sourced

9. ZeroGUI

- Title: ZeroGUI: Automating Online GUI Learning at Zero Human Cost
- Date: 2025/05/29
- Input: Screenshot
- Method:
    - online learning framework
    - VLM-based automatic task generation (GPT-4o) + VLM-based automatic reward estimation (Qwen2.5-VL-32B) + two-stage online reinforcement learning to continuously interact with and learn from GUI environments
- Base Models: UI-TARS-7B-DPO and Aguvis-7B as base models for training

- Platforms: desktop(OSWorld) + mobile(AndroidLab)
- Performance Reported: Improves Performance Reported of UI-TARS and AGUVIS
  - OSWorld:
    - GPT-4o: 5.0
    - Gemini-Pro-1.5: 5.4
    - Claude Computer-Use: 14.9
    - OpenAI Operator: 19.7
    - CogAgent-9B: 8.1
    - Aguvis-72B: 10.3
    - UI-TARS-72B-DPO: 22.7
    - Aguvis-7B: 3.0
    - Aguvis-7B + ZeroGUI: 4.9
    - UI-TARS-7B-DPO: 17.7
    - UI-TARS-7B-DPO + ZeroGUI: 20.2
  - AndroidLab:
    - UI-TARS-7B-DPO:45.7
    - UI-TARS-7B-DPO+ZeroGUI:47.5
- Availability: Training code samples open sourced

10. GUI-Robust

- Title: GUI-Robust: A Comprehensive Dataset for Testing GUI Agent Robustness in Real-World Anomalies
- Date: 2025/06/17
- Dataset: a dataset designed for GUI agent evaluation, and a semi-automated dataset construction paradigm via RPA tools, YOLOv8 for UI element detection and Qwen2.5-VL to generate
- Platforms: Desktop (Apps and Web)
- Availability: Dataset open sourced

11. AgentCPM-GUI

- Title: AgentCPM-GUI: Building Mobile-Use Agents with Reinforcement Fine-Tuning
- Date: 2025/06/17
- Dataset: CAGUI benchmark — 55K trajectories, 470K steps from Chinese Android apps
- Base Model: Fine-tuned from MiniCPM-V-2_6 (8B)
- Method:
  - Grounding-aware pretraining
  - Supervised fine-tuning (SFT)
  - Reinforcement fine-tuning (GRPO)
- Features: Compact action space; supports inference via vLLM
- Platforms: Mobile (Android)
- Availability: SFT and RFT code open sourced; pretraining code not released; model open sourced; dataset proprietary

12. GLM-4.1V-Thinking

- Title: GLM-4.1V-Thinking: Towards Versatile Multimodal Reasoning with Scalable Reinforcement Learning
- Date: 2025/07/02

- Input: Screenshot
- Base Model: GLM-4.1V-9B-Thinking
- Performance Reported: Outperforms Qwen2.5-VL-7B and GPT-4o (2024-11-20) in GUI tasks
  - OSWorld:
    - GLM-4.1V 9B-Thinking: 14.9
    - Qwen2.5-VL 7B: 1.9
    - InternVL3 9B: 1.4
    - Kimi-VL A3B-Thinking: 8.2
    - MiMo-VL 7B-RL: 1.9
    - Qwen2.5-VL 72B: 8.8
    - GPT-4o (2024-11-20): 5.0
  - AndroidWorld:
    - GLM-4.1V 9B-Thinking: 41.7
    - Qwen2.5-VL 7B: 27.6
    - InternVL3 9B: 1.9
    - Kimi-VL A3B-Thinking: -
    - MiMo-VL 7B-RL: 10.8
    - Qwen2.5-VL 72B: 35.0
    - GPT-4o (2024-11-20): 34.5
- Method: Reinforcement Learning with Curriculum Sampling (RLCS)
- Platforms: Mobile, Desktop
- Availability:
  - Model: MIT License
  - Code: Apache License 2.0
  - SFT Support: LLaMA-Factory