Video2GUI: Synthesizing Large-Scale Interaction Trajectories for Generalized GUI Agent Pretraining

Weimin Xiong, Shuhao Gu, Bowen Ye, Zihao Yue, Lei Li, Feifan Song, Sujian Li, Hao Tian

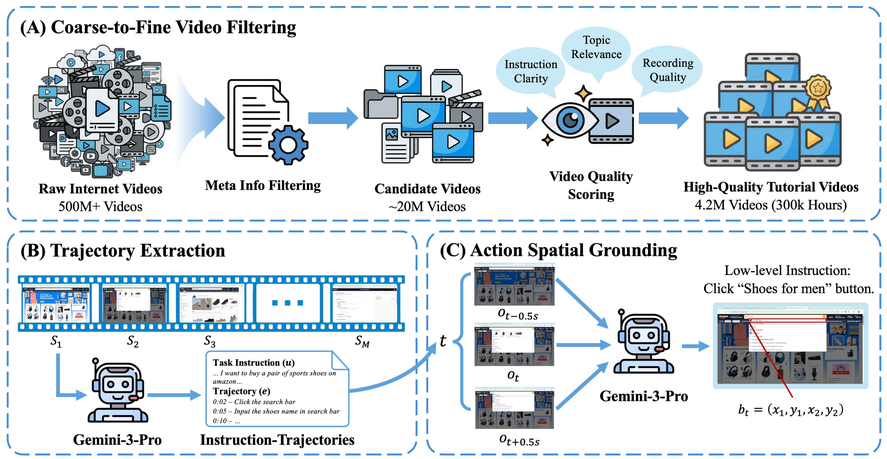
ICML International Conference On Machine Learning

Introduction

- **TL;DR:** We present **Video2GUI**, a fully automated framework that extracts grounded GUI interaction trajectories directly from unlabeled Internet videos.
- **Motivation:** Generalized GUI agents depend on large-scale, diverse trajectory data spanning real-world applications. But existing datasets rely on costly manual annotation and are confined to narrow domains or single platforms — a key bottleneck for agent generalization.
- **Key Challenges:** Internet videos are a rich, free source of real GUI usage — but hard to exploit:
  - Filtering: identifying high-quality GUI tutorials among billions of noisy videos
  - Annotation: raw videos lack interaction labels and precise grounding coordinates
- **Here's what we get: WildGUI** — the largest open-source GUI pre-training dataset across web / mobile / desktop platform. Pre-training Qwen2.5-VL & Mimo-VL yields 5–20% gains, matching or surpassing SOTA.

Dataset	Platform			Scale & Statistics				Inst. Level
	Website	Mobile	Desktop	Environments	Instructions	Images	Turns	
MiniWoB++ (Liu et al., 2018)	✓	✓	✗	114	100	17,971	3.6	Low-level
MIND2WEB (Deng et al., 2023)	✓	✗	✗	137	2,350	2,350	7.3	High-level
AITW (Rawles et al., 2023)	✗	✓	✗	357	30,378	715,142	6.5	High & low
AndroidControl (Li et al., 2024)	✗	✓	✗	833	14,538	15,283	4.8	High & low
GUI-World (Chen et al., 2024)	✓	✓	✓	-	12,379	83,176	6.7	High-level
GUI-Odessey (Lu et al., 2025c)	✗	✓	✗	201	7,735	118,791	15.4	High-level
GUI-Act (Chen et al., 2025)	✓	✗	✗	50	67k	13k	7.9	Low-level
GUI-Net (Zhang et al., 2025a)	✓	✓	✓	280	1M	1M	4.7	High-level
MONDAY (Jang et al., 2025)	✗	✓	✗	-	20K	313k	15.7	High-level
GUI-360° (Mu et al., 2025)	✗	✗	✓	3	13,750	105,368	7.6	High-level
WildGUI (Ours)	✓	✓	✓	1,500+	12.7M	124.5M	9.7	High & low

Method



Experiments

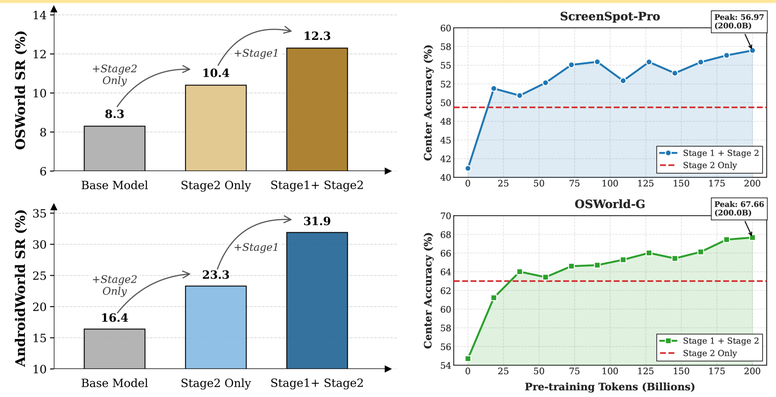
GUI Grounding Evaluation

Agent Model	ScreenSpot-Pro			OSWorld-G				
	Text	Icon	Avg	Text Match.	Elem. Rec.	Layout Und.	Fine-grained	Avg
Proprietary Models								
Gemini-2.5-Pro	-	-	11.4	59.8	45.5	49.0	33.6	45.2
Seed1.5-VL	-	-	60.9	73.9	66.7	69.6	47.0	62.9
Open-Source Models								
Qwen3-VL-2B* (Bai et al., 2025a)	56.1	18.9	41.9	61.7	45.8	54.2	39.6	45.9
GTA1-7B (Yang et al., 2025)	65.5	25.3	50.1	42.1	65.7	62.7	56.1	55.1
UI-Venus-7B (Gu et al., 2025)	67.1	24.3	50.8	74.6	60.5	61.5	45.5	58.8
OpenCUA-7B (Wang et al., 2025)	-	-	50.0	-	-	-	-	55.3
GUI-Owl-7B (Ye et al., 2025)	69.4	31.5	54.9	64.8	63.6	61.3	41.0	55.9
Qwen3-VL-8B* (Bai et al., 2025a)	67.6	21.3	49.9	69.0	55.5	59.7	47.7	54.8
Qwen3-VL-32B* (Bai et al., 2025a)	73.4	25.0	54.9	72.8	63.3	66.4	51.7	60.6
UI-TARS-72B (Qin et al., 2025)	50.9	17.5	38.1	69.4	60.6	62.9	45.6	57.1
Effectiveness of WildGUI (Ours)								
Qwen2.5-VL-7B (Bai et al., 2025b)*	-	-	26.8	41.4	28.8	34.8	13.4	27.3
+ WildGUI	57.0	17.6	41.9 (↑15.1)	70.0	54.6	57.7	46.2	53.7 (↑26.4)
Mimo-VL-7B (Xiaomi, 2025)	55.7	18.4	41.2	65.0	59.2	59.0	40.2	54.7
+ WildGUI	70.1	33.6	**56.9** (↑15.7)	80.8	68.3	71.1	61.4	**67.6** (↑12.9)

Offline GUI Agent Evaluation

Models	AndroidControl-Low		AndroidControl-High		CAGUI	
	Type Acc.	Step SR	Type Acc.	Step SR	Type Acc.	Step SR
Closed-source Models						
GPT-4o (Hurst et al., 2024)	74.3	19.4	66.3	20.8	3.7	3.7
Open-source Models						
OS-Genesis-7B (Sun et al., 2025)	90.7	74.2	65.9	44.4	38.1	14.5
OS-Atlas-7B (Wu et al., 2024)	73.0	67.3	70.4	56.5	81.5	55.9
Aguvis-7B (Xu et al., 2024)	93.9	89.4	65.6	54.2	67.4	38.2
UI-TARS-7B (Qin et al., 2025)	98.0	90.8	83.7	72.5	88.6	70.3
Effectiveness of WildGUI (Ours)						
Qwen2.5-VL-7B* (Bai et al., 2025b)	94.1	85.0	75.1	62.9	74.2	55.2
+ WildGUI	94.9 (↑0.8)	90.3 (↑5.3)	74.6	64.5 (↑1.6)	88.3 (↑14.1)	65.4 (↑10.2)
Mimo-VL-7B (Xiaomi, 2025)	92.9	87.9	76.3	65.6	82.2	63.4
+ WildGUI	**95.5** (↑2.6)	**91.8** (↑3.9)	**80.6** (↑4.3)	**71.4** (↑5.8)	**90.3** (↑8.1)	**71.0** (↑7.6)

Online Evaluation & Scaling Effects

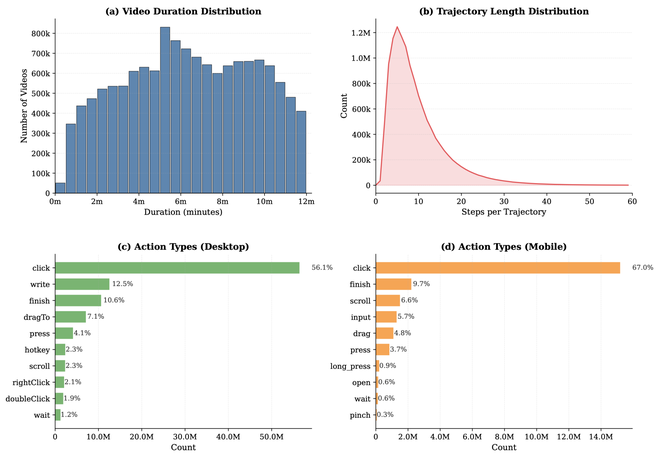


Discussion

D1: Ablation Studies

Setting	ScreenSpot-Pro	CAGUI	AndroidWorld
Ours	**56.9**	**71.0**	**31.9**
w/o $\mathcal{L}_{ground}$	49.8	69.8	28.4
w/o $\mathcal{L}_{action}$	50.5	65.3	27.6
w/o $\mathcal{L}_{traj}$	54.6	70.2	24.1
w/o Stage 1	49.3	64.2	23.3
w/o Stage 2	28.2	45.7	6.0

D2: Dataset statistics of WildGUI



Contact Info

Paper

Dataset

WeChat

- Graduating in 2028; seeking industry/academic opportunities
- Email: wmxiong@pku.edu.cn; lisujian@pku.edu.cn;
- Twitter: @WeiminXiong