MONDAY: Scalable Video-to-Dataset Generation for Cross-Platform Mobile Agents

Yunseok Jang[1*], Yeda Song[1*], Sungryull Sohn[2], Lajanugen Logeswaran[2], Tiange Luo[1], Dong-Ki Kim[2], Kyunghoon Bae[2], Honglak Lee[1,2]

[1]University of Michigan [2]LG AI Research *Equal contribution

Data & Paper

CVPR Nashville JUNE 11-15, 2025

Motivation

Existing 📱 **mobile GUI agent** datasets & environments:

⚠️ Mostly single OS (Android)

⚠️ Low diversity in UI and user configs

⚠️ Expensive to collect and annotate

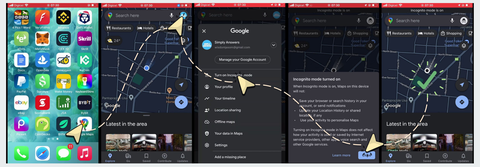
Solution: **Transform internet videos into GUI agent data**

New GUI Dataset: MONDAY

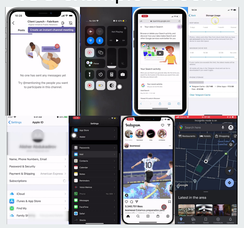
(**M**obile **O**S **N**avigation Task **D**ataset for **A**gents from **Y**ouTube)

Dataset	# Videos	# Frames	Real-world	Data Access	Code Access	Platforms	Method
RICO	✗	72K	✗	✓	✗	Android	Manual
AitW	✗	5.7M	✗	✓	✗	Android	Manual
MM-Navigator	✗	110	✓	✗	✓	iOS + Android	Manual
Chen et al.	128	1K	✓	✗	✗	iOS + Android	Manual
Video2Action	6K	30K	✓	✗	✗	Android	Semi-Auto
MONDAY (Ours)	20K	313K	✓	✓	✓	iOS + Android	Automated

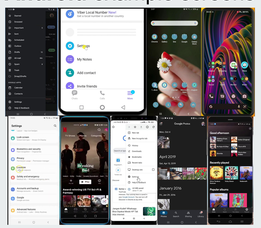
Task: “How to turn on incognito mode on Google Maps?”



iOS example screens



Android example screens



We thank the many prior works that made the development of MONDAY possible; see the paper for a complete list.

Liu et al., *Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection*, in ECCV, 2024

Yang et al., *Set-of-Mark Prompting Unleashes Extraordinary Visual Grounding in GPT-4V*, arXiv:2310.11441

Rawles et al., *Android in the Wild: A Large-Scale Dataset for Android Device Control*, in NeurIPS, 2023

Chai et al., *AMEX: Android Multi-annotation Expo Dataset for Mobile GUI Agents*, arXiv:2407.17490

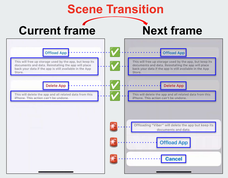
Diverse, large-scale, real-world GUI agent dataset from internet videos, via fully automated pipeline

Video Collection

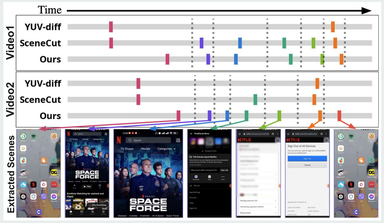
1. Collect **how-to posts** for mobile devices (C4, Dolma)
2. Filter mobile agent tasks; **extract titles** (GPT 3.5 Turbo)
3. **Search videos** 

4. **Filter videos** 129K → 20K (15.5%)

Scene Transition Detection



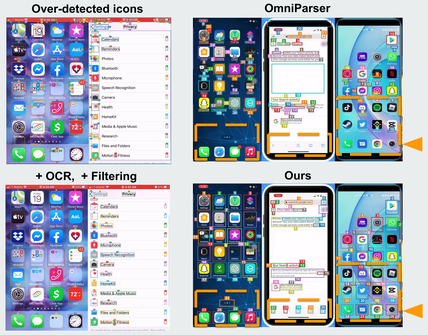
Scene Transition = **OCR Box Diff** > Threshold



	YUV-diff	SceneDetect	OCR-based (Ours)
F1-score (%)	70.86	82.27	**95.04**

UI Element Detection

1. Over-detect **icons** with GroundingDINO
2. **OCR** bounding box detection
3. Heuristic **rule-based** filtering



	OmniParser	Ours
Hit Ratio (%)	91.83	**99.87**

3-Step Action Identification

1. Scene summary

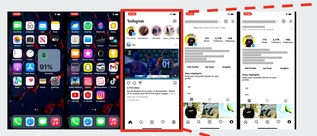
Describe each frame and its visible UI elements.

2. Initial action identification

Identify the action and potential interaction area.

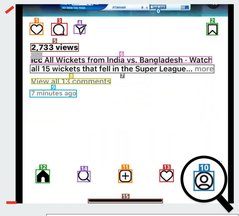
3. Refined action identification

Re-analyze the zoomed-in area to precisely localize the action.



Current Scene

Higher Resolution



narration → GPT-4o

Scene summary:
"-2": Home screen …,
"-1": Second home screen …,
"0": Instagram home feed …,
"+1": Instagram profile page …,
"+2": Profile dashboard …

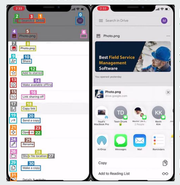
scene summary → GPT-4o

Action type: Touch
Bbox: ID=40, Zone 5
Detail: Profile icon on the bottom-right corner

GPT-4o

Action type: Touch
Bbox: ID=10, [0.881, 0.937]
Detail: Profile icon on the bottom-right corner

Ablation study



Identified Actions

Multi-image 3-step (Ours): Touch, ‘Send a copy’ option, box 20

2-step: Touch, ‘Send a copy’ option, box 20

1-step: Scroll, up

No narrations: Touch, ‘Share’ option, box 10

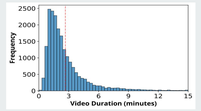
First-step w/ single-image: Touch, ‘Share’ option, box 10

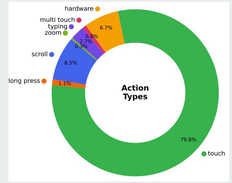
Method	All (%)	Touch (%)
Multi-image 3-step (**Ours**)	**80.90**	**91.84**
Number of steps:		
2-step	79.43	89.97
1-step	70.63	74.67
Missing context:		
No narrations	78.20	87.64
First-step w/ single-image	77.22	89.30

Accuracy (%)

MONDAY: Data Statistics

	iOS	Android	Total
Train	9,755	9,970	19,725
Val	246	249	495
Test	50	50	100
Total	**10,051**	**10,269**	**20,320**





Apps Count		Value
Total		**2,479**
Type	OS native	37.6%
	Third-party	62.4%
Top 5 Apps	Instagram	3.7%
	Facebook	2.6%
	YouTube	2.1%
	Twitter	1.9%
	WhatsApp	1.8%

Agent Evaluation Results

Action prediction: $P(a_t \mid \text{instruction}, (a_{t-i})_{i=1}^{h}, o_t)$

① Base models → Finetune on AitW/AMEX

② Base models → **Finetune on MONDAY** → Finetune on AitW/AMEX

Finetuned model	Test set			
	AitW	AMEX	MONDAY	Windows Mobile
AitW-finetuned from:				
SeeClick	66.98	47.23	40.66	38.54
SeeClick-**MONDAY**	**68.47**	**47.76**	**63.39**	**51.71**
AMEX-finetuned from:				
SeeClick	37.08	**68.19**	44.23	43.17
SeeClick-**MONDAY**	**40.19**	66.13	**63.39**	**55.37**
AitW-finetuned from:				
Llama-3.2	58.96	43.74	39.8	26.83
Llama-3.2-**MONDAY**	**67.38**	**55.96**	**57.99**	**50.24**
AMEX-finetuned from:				
Llama-3.2	29.81	61.3	40.17	28.29
Llama-3.2-**MONDAY**	**42.96**	**72.36**	**58.35**	**51.46**

Step Success Rate (%)

✔️ Cross-platform generalization

✔️ Out-of-distribution generalization

Step SR on Windows Mobile