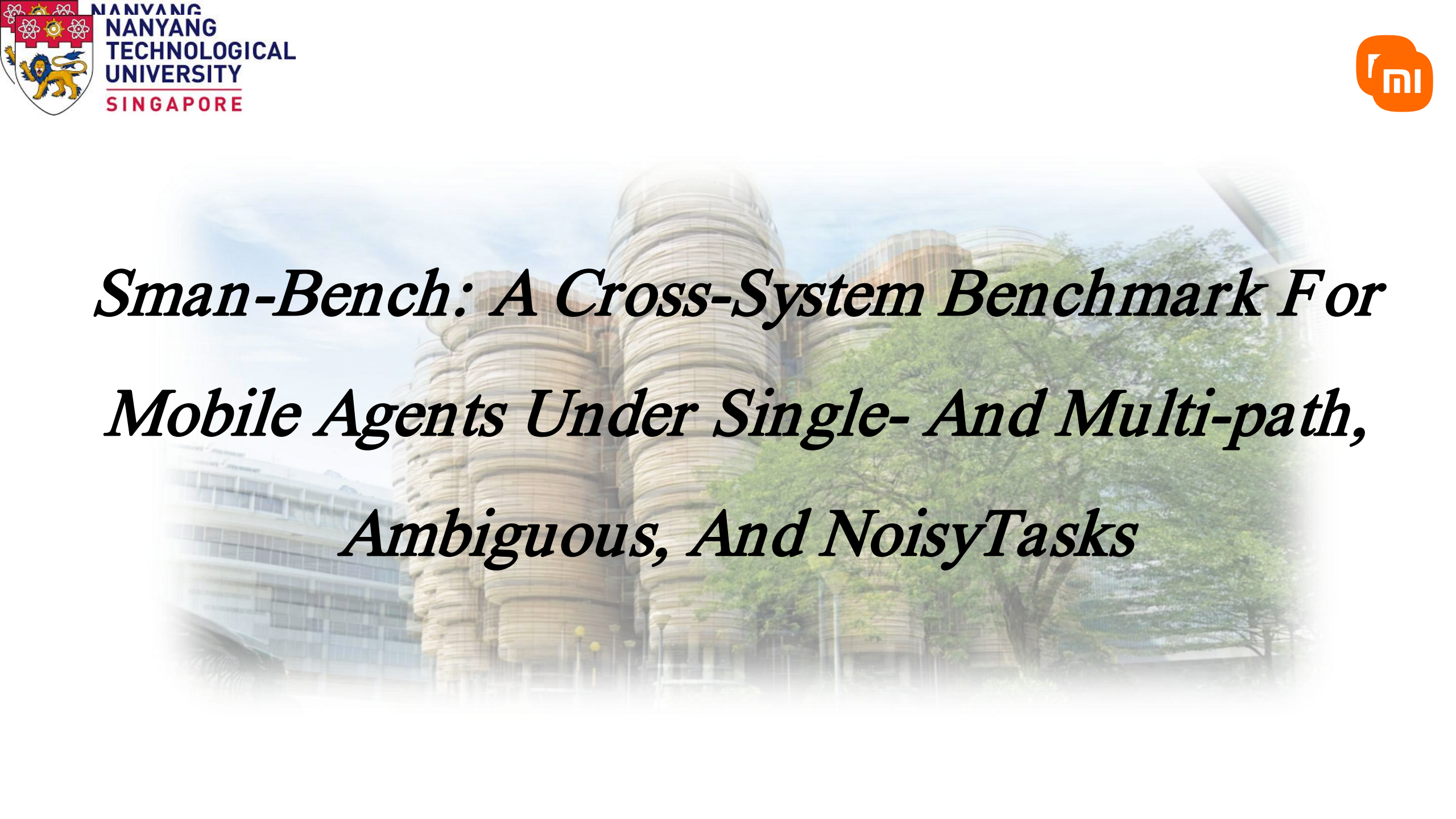


Sman-Bench: A Cross-System Benchmark For Mobile Agents Under Single- And Multi-path, Ambiguous, And NoisyTasks

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Introduction

- ***Reward Instability***: *Existing online benchmarks fail to get stable critical reward signals under dynamic environmental changes.*

- ***Single-path Limitation***: *Offline benchmarks evaluate agents through single-path trajectories, which contradicts the inherently multi-solution characteristics of GUI tasks.*

- ***Noise Neglect & Proactive Interaction***: *Existing benchmarks neglect the influence of noise components and the need for proactive interaction under ambiguous instructions.*

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Contribution

- *We construct a cross-system benchmark named SMAN-Bench and propose a slot-based instruction generation method named GIAS.*
- *We propose an offline multi-path evaluation method and leverage slot-based key node annotations to enable stable assessment of step rewards.*
- *We introduce SMAN-Bench-Noisy to support realistic noisy evaluation by collecting data from noisy apps, enabling robust assessment under challenging environments.*
- *We propose SMAN-Bench-Ambiguous to facilitate active interactive evaluation, where agents are allowed to ask clarification questions during execution.*

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Statistics

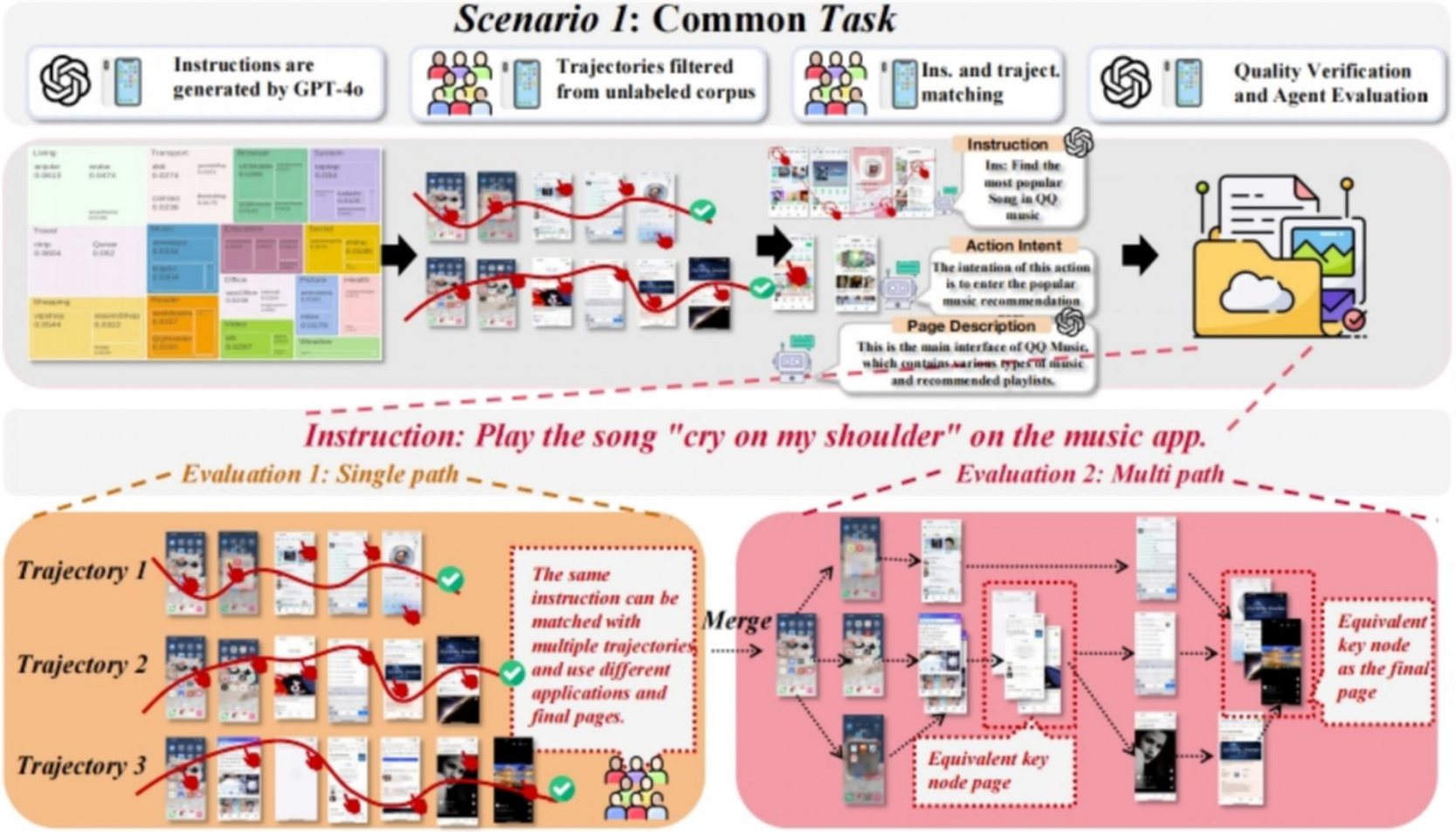
- ***Broad App Coverage***: *The benchmark comprises 15 categories and 49 widely used apps.*
- ***Common Task Scale***: *The common split includes 12,854 instructions and 800 templates generated by GIAS.*
- ***Balanced Complexity***: *There are 9,620 simple tasks with an average of 5.62 steps and 3,234 complex tasks with an average of 8.21 steps.*

Table 1: Comparison of SMAN-Bench to other benchmarks.

Benchmarks	# Inst.	Language	# Avg Steps	# Screen-shots	Path	Online Environ.?	Ambi. Noise	Interac. Inst.
PIXELHELP	187	EN	4.2	~800	Single	✗	✗	✗
MOTIF	480	EN	4.5	~21K	Single	✗	✗	✗
AMEX	341	EN&CN	12.8	~104K	Single	✗	✗	✗
SCREENSPOT	~1,200	EN&CN	1	~600	Dot	✗	✗	✗
MOBILEAIBENCH	*	EN	*	*	Dot	✗	*	✗
AGENTBENCH	100	EN	20	~2k	Multiple	✓	✗	✗
GUI ODYSSEY	7,735	EN	15.4	*	Single	✗	✗	✗
MOBILE-BENCH	832	CN	*	14,144	Multiple	✓	✗	✗
MVISU-BENCH	404	EN&CN	*	*	Multiple	✓	✗	✓
SPA-BENCH	340	EN&CN	*	*	Multiple	✓	✗	✗
ANDROIDLAB	10.5k	EN	8.98	94.3k	Multiple	✓	✗	✗
ANDROIDWORLD	116	EN	*	*	Multiple	✓	✗	✓
SMAN-BENCH	12,856	EN&CN	7.28	~48k	Both	✗	✓	✓

Common Task

- ***Automated Instruction Generation (GIAS):*** *Uses intent inference and extracts slot information from GUI changes to transform unlabeled action sequences into natural language instructions.*
- ***Slot Matching Mechanism:*** *Employs a slot-based template filling mechanism, allowing one template to match multiple trajectories that share key nodes.*



Noisy & Ambiguous Tasks

- ***Noisy Task:*** *Collects data from apps with substantial ads and pop-ups, including static pop-ups, dynamic video ads, and redirecting links, to evaluate noise robustness.*
- ***Ambiguous Task:*** *Progressively simplifies full instructions into ambiguous ones by removing slots, assigning preset Q&As to corresponding GUIs to evaluate proactive interaction.*

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Multi-path Evaluation

- ***Graph Structure Exploration**: Allows the agent to freely explore within the pre-executed graph corpus, provided the maximum step limit is not exceeded.*
- ***Node Merging**: Merges discrete single trajectories into a unified graph using action space, pixel differences (BM25), and consistency in XML button values.*
- ***Evaluation Advantages**: Combines the advantages of online and offline environments, allowing for a fairer outcome-based comparison across different agents.*

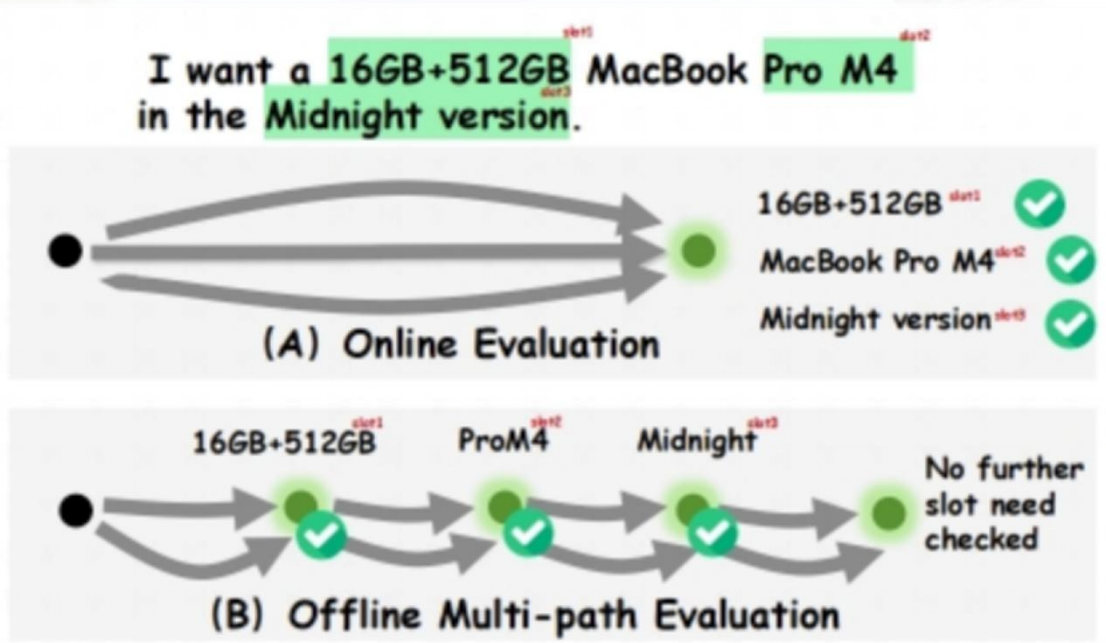


Figure 3: Unlike Online Evaluation, offline multi-path evaluation checks both process and final GUI as reward signals.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Main Results

Models	Cate.	Common-Simple Type↑/SE↓	Common-Simple Step. Acc	Common-Simple SR	Common-Complex Type↑/SE↓	Common-Complex Step. Acc	Common-Complex SR	Noisy Data Type↑/SE↓	Noisy Data Step. Acc	Noisy Data SR	Ambiguous Data Type↑/SE↓	Ambiguous Data Step. Acc	Ambiguous Data SR
Single-Agent Framework: AppAgent-v1													
InternVL2-40B	Single	82.4	35.7	1.0	85.1	38.4	1.5	60.4	12.7	0.0	76.7	29.0	1.0
	Multi	5.8	43.0	10.1	4.4	46.0	6.5	-	-	-	-	-	-
LLAVA-72B-NEXT	Single	93.2	7.4	0.0	87.7	7.8	0.0	70.4	2.7	0.0	82.3	4.8	0.0
	Multi	6.2	1.9	0.0	4.5	5.1	0.0	-	-	-	-	-	-
Qwen2-VL-72B	Single	95.2	60.3	21.1	93.5	53.8	5.0	78.0	24.4	3.0	91.2	43.5	8.0
	Multi	5.2	62.8	20.6	4.4	58.9	7.0	-	-	-	-	-	-
Qwen-VL-Max	Single	94.7	58.6	20.5	91.2	54.7	7.5	77.1	24.3	3.0	90.3	48.5	9.0
	Multi	5.9	67.6	12.6	4.3	63.1	9.6	-	-	-	-	-	-
Llama3.2-VL-90B	Single	86.4	22.4	2.6	87.0	24.3	1.0	69.7	11.2	0.0	85.4	15.5	0.0
GPT-4v	Single	91.2	24.0	6.0	90.8	25.2	1.0	72.7	17.4	0.0	88.6	20.5	1.0
	Multi	6.1	29.7	3.0	4.5	29.4	1.5	-	-	-	-	-	-
GPT-4o	Single	80.4	57.6	18.5	79.2	50.6	11.5	72.3	18.2	1.0	94.7	33.9	6.0
	Multi	5.3	61.8	19.8	4.4	61.7	16.5	-	-	-	-	-	-
Multi-Agents Framework: MobileAgent-v2													
InternVL2-40B	Single	84.5	19.3	0.0	80.7	26.4	0.5	64.3	9.1	0.0	76.0	16.3	0.0
	Multi	6.0	27.6	3.5	4.4	32.6	3.5	-	-	-	-	-	-
Qwen2-VL-72B	Single	91.5	50.5	13.0	91.6	49.0	4.5	75.8	20.7	1.0	86.2	40.8	7.0
	Multi	5.4	54.9	15.1	4.4	58.6	8.0	-	-	-	-	-	-
Qwen-VL-Max	Single	74.2	17.0	3.0	68.8	12.3	2.0	66.5	4.2	0.0	66.6	9.0	0.0
	Multi	5.4	29.6	4.5	4.3	24.8	3.0	-	-	-	-	-	-
Llama3.2-VL-90B	Single	62.4	16.6	1.0	67.0	17.5	0.0	63.7	9.7	0.0	64.3	8.3	0.0
GPT-4v	Single	90.8	22.9	3.8	90.6	28.3	0.5	62.5	12.6	0.0	91.0	15.6	0.0
	Multi	6.0	17.8	5.4	4.5	11.8	0.0	-	-	-	-	-	-
GPT-4o	Single	91.9	53.5	13.5	92.3	50.5	17.0	77.1	25.5	2.0	91.6	39.7	12.0
	Multi	4.9	57.6	25.5	4.2	56.3	17.5	-	-	-	-	-	-
Multi-Agents Framework: MobileAgent-E													
InternVL2-40B	Single	87.2	37.4	2.0	93.4	39.1	2.5	80.3	23.1	4.0	87.0	32.6	7.0
	Multi	5.5	46.2	9.5	4.2	46.7	8.5	-	-	-	-	-	-
Qwen2-VL-72B	Single	92.3	62.5	23.0	93.2	61.4	15.0	88.9	32.7	7.0	91.0	58.8	15.0
	Multi	4.8	60.2	18.4	4.2	63.8	12.0	-	-	-	-	-	-
Qwen-VL-Max	Single	90.6	71.5	32.5	94.8	66.3	25.5	86.2	61.6	21.0	96.3	62.5	29.0
	Multi	3.7	77.1	29.5	3.1	78.8	33.5	-	-	-	-	-	-
Llama3.2-VL-90B	Single	78.2	29.8	4.0	87.2	32.4	6.5	80.7	22.1	3.0	79.8	23.4	2.0
GPT-4o	Single	94.8	70.3	27.5	96.4	68.4	19.0	81.2	54.8	14.0	92.6	53.4	24.0
	Multi	3.9	77.7	30.5	3.8	70.4	26.5	-	-	-	-	-	-

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Main Results

Models	Cate.	Common-Simple			Common-Complex			Noisy Data			Ambiguous Data		
		Type	Step. Acc	SR	Type	Step. Acc	SR	Type	Step. Acc	SR	Type	Step. Acc	SR
Continuous Pre-training Mobile Agents													
CogAgent-18B	Single	75.6	20.9	13.0	62.3	20.8	6.0	57.2	16.2	1.0	72.4	30.6	11.0
UGround-7B	Single	73.0	39.5	17.0	73.8	36.0	10.0	71.2	32.8	2.0	79.9	47.0	20.0
UI-Tars-7B-dpo	Single	75.3	41.8	19.0	75.9	37.8	12.0	73.2	35.6	4.0	81.8	49.4	23.0
OS-Atlas-7B-pro	Single	82.1	51.5	28.0	83.5	50.6	18.0	81.2	45.3	2.0	86.4	52.3	24.0
Kimi-VL-A3B	Single	75.1	21.7	12.0	61.8	21.5	6.5	56.9	15.8	1.0	71.7	29.9	11.0
DeepSeek-VL2	Single	72.6	38.8	16.0	73.1	35.2	9.5	70.5	32.1	2.0	79.2	46.1	19.0
UI-Tars-72B-dpo	Single	94.3	64.2	32.0	96.0	63.5	24.0	94.5	59.8	7.0	92.5	66.0	30.0
GUI-OWL-7B	Single	94.7	71.2	37.5	96.0	68.4	28.5	90.8	52.4	6.0	93.8	67.2	31.0
UI-TARS-1.5-7B	Single	98.2	72.2	**39.0**	97.1	77.5	38.5	**98.0**	67.3	15.0	99.0	78.4	42.0
OpenCUA-32B	Single	98.0	73.1	**39.0**	97.4	76.2	38.0	96.0	65.8	13.5	98.2	**79.2**	**43.0**
RL-based Mobile Agents													
UI-R1-3B	Single	76.5	42.7	18.5	74.8	39.1	10.5	72.6	33.9	6.0	80.9	47.8	21.5
GUI-R1-3B	Single	77.2	40.9	20.0	76.4	38.6	11.0	74.1	36.8	5.5	82.1	48.6	22.5
GUI-G1-3B	Single	82.4	50.8	25.5	84.1	52.3	19.0	80.6	48.7	16.5	85.3	54.9	23.0
GUI-R1-7B	Single	92.8	62.7	30.5	94.2	61.9	22.5	92.7	58.1	5.5	90.9	64.3	28.5
UI-S1-7B	Single	95.7	65.8	33.0	97.5	65.1	25.5	96.1	61.5	8.5	94.2	67.5	31.5
Reasoning Mobile Agents													
GLM-4.1v-Thinking	Single	78.4	42.8	17.5	80.7	44.6	17.0	71.3	30.2	2.5	77.8	44.9	18.0
Qwen-QVQ-plus	Single	90.7	48.8	24.5	89.9	52.8	18.0	88.1	44.4	6.5	92.2	64.4	29.0
OpenAI o3-2025-04-16	Single	94.2	68.1	33.0	95.5	58.2	21.5	90.7	52.2	9.0	94.9	72.3	33.5
Claude 3.7 Sonnet	Single	98.4	74.4	38.0	98.0	76.1	38.5	95.5	59.2	10.0	**99.0**	77.3	41.0
Doubao-1.5-Thinking-pro	Single	98.2	75.8	**39.0**	98.1	**77.3**	38.5	95.9	60.1	14.0	**99.0**	78.0	41.5
Claude 4.5 Sonnet	Single	**98.9**	**76.2**	**39.0**	**98.5**	77.1	**39.0**	**98.5**	**69.2**	**15.5**	**99.0**	78.3	**43.0**

Conclusion

- *Framework Comparison: AppAgent-v1 outperforms in single-path evaluations, whereas Mobile-Agent-v2 and Mobile-Agent-E (with dynamic knowledge injection) excel in multi-path scenarios.*
- *Noise Vulnerability: All VLMs exhibit a declining trend in Step Accuracy under noisy data, indicating that open-source agents fail to learn advertisement features and lack generalization.*
- *Proactive Interaction Benefits: Ablation results demonstrate the active interaction module helps agents ignore irrelevant content, with some VLMs like Llama3.2-VL-90B showing up to a 17.5% performance improvement.*