VenusBench-Mobile

A Challenging and User-Centric Benchmark for Mobile GUI Agents with Capability Diagnostics

Yichen Gong* **Zhuohan Cai*** Sunhao Dai Yuqi Zhou Zhangxuan Gu Changhua Meng Shuheng Shen

Presenter: **Zhuohan Cai**

Department of Computer Science and Technology, Tsinghua University

Venus Team, Ant Group

Tsinghua University

Ant Group

ICML 2026

ICML 2026 Oral

Roadmap: From User Needs to Failure Diagnosis

Question 1: What should we evaluate?

User-intent-driven tasks that reflect real mobile usage.

Question 2: How should we evaluate?

Capability-level diagnostics instead of only final success rates.

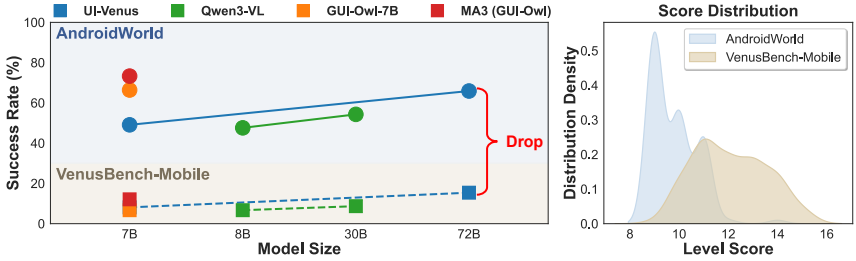
Question 3: What do current agents fail at?

Perception, memory, and robustness under realistic variations.

Takeaway

VenusBench-Mobile exposes a large gap between benchmark success and real-world reliability.

Existing Benchmarks Hide the Deployment Gap



Observed mismatch

- Agents that look strong on AndroidWorld drop sharply on VenusBench-Mobile.
- The gap comes from higher user-intent and capability requirements.

Core concern

High benchmark scores do not necessarily imply reliable mobile assistance.

User Intent Should Be the Evaluation Unit

Prior view: app-centric

- Choose apps first.
- Create tasks around app functionalities.
- Measure completion with coarse success rate.

Our view: user-centric

- Start from real user needs.
- Treat apps as tools inside a phone environment.
- Diagnose which capability caused failure.

A general mobile assistant should be evaluated by user intent, not by isolated app functions.

User Intent Defines the Task Space

FA Function assistance

CF Conflict handling

VA Vague instructions

MR Multi-round interaction

GSA GUI state awareness

GUIM Visual GUI manipulation

HGB Hard GUI browsing

NR Noise resistance

BC Browsecomp-like tasks

SE Stability evaluation

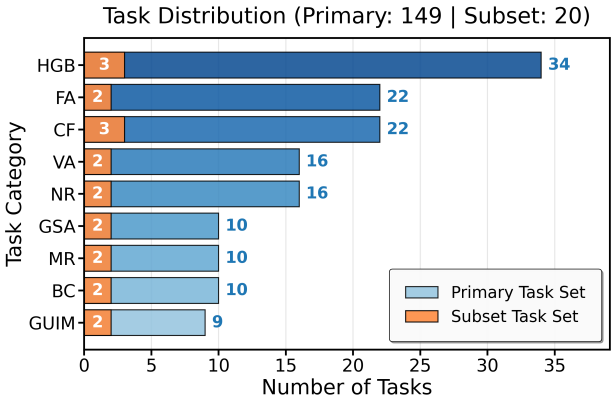
Design choice

Start from user needs, then instantiate tasks in Android apps.

What becomes visible

- Ambiguous or impossible requests
- Dynamic GUI states and multi-turn goals
- Visual editing, browsing, and disruptions

149 Tasks Stress Real Mobile Assistance



Dataset scale

- 27 open-source Android apps
- 149 primary tasks
- 80 additional stability variants

Balanced stress testing

HGB, FA, CF, VA, NR and other categories expose different real-world failure modes.

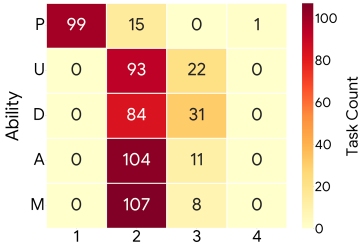
PUDAM Turns Success Rate into Capability Requirements

PUDAM

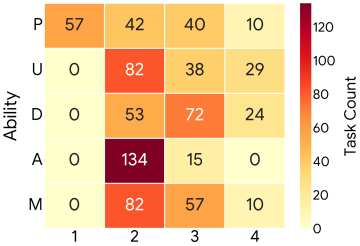
- **P**: Perception
- **U**: Understanding
- **D**: Decision
- **A**: Action
- **M**: Memory

Key difference

Each task is annotated by required capability and four proficiency levels.



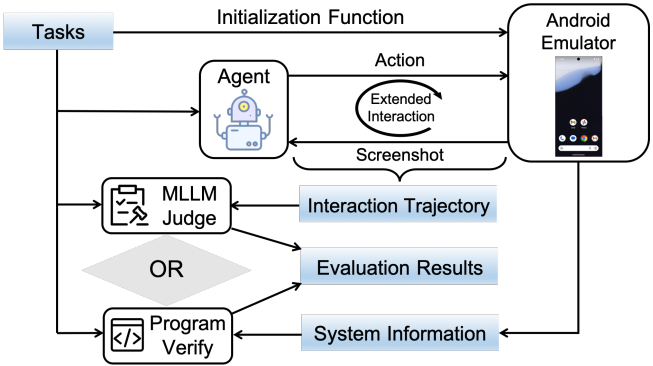
AndroidWorld: lower-level concentration



VenusBench-Mobile: broader high-level coverage

The same final failure can now be mapped to different missing capabilities.

Online Infrastructure Makes Realistic Failures Verifiable



Closed-loop Android execution with hybrid task verification.

Execution loop

Observe screenshots, act in the emulator, then receive state updates.

Hybrid verification

Program checks verify states; MLLM judge handles semantic or visual goals.

Extended scenarios

Multi-round interaction and dynamic noise expose reproducible failures.

Prior Online Benchmarks Miss Key User-Centric Scenarios

Benchmark	# Apps	# Tasks	Verif.	Cost	Task Categories									
					FA	CF	VA	MR	GSA	GUIM	HGB	NR	BC	ST
LearnGUI	20	101	✓								✓			
MMBench-GUI	-	146	✓											
UI-NEXUS	20	100	✓	✓	✓									
MVISU	137	404		✓		✓	✓				✓			
AndroidWorld	20	116	✓	✓							✓			✓
AndroidLab	9	138	✓	✓										
MobileAgentBench	10	100	✓	✓										
AndroidDaily	48	235									✓			
SPABench	66	340		✓										
MobileWorld (GUI)	20	161	✓			✓	✓	✓						
VenusBench-Mobile (Ours)	27	149(+80)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Blind spots

Prior benchmarks rarely cover conflict handling, vague instructions, GUI state awareness, visual manipulation, and stability together.

Main Result: Current Agents Remain Far from Reliable

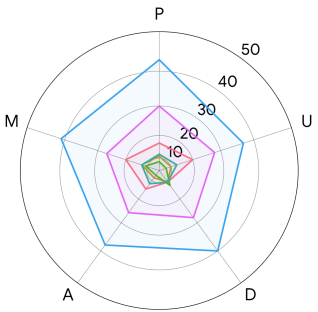
Model	FA	CF	VA	MR	GSA	GUIM	HGB	NR	BC	Total
Open-source										
UI-Venus-72B	22.7	4.6	12.5	0.0	10.0	0.0	17.7	50.0	0.0	15.4
UI-Venus-7B	13.6	4.6	25.0	0.0	10.0	0.0	0.0	18.8	0.0	8.1
Qwen3-VL-30B-A3B	22.7	4.6	18.8	0.0	0.0	0.0	5.9	6.3	10.0	8.7
Qwen3-VL-8B	18.2	4.6	18.8	0.0	0.0	0.0	0.0	6.3	10.0	6.7
GUI-Owl-7B	13.6	0.0	18.8	0.0	0.0	11.1	2.9	12.5	0.0	6.7
MA3 (GUI-Owl-7B)	18.2	9.1	6.3	0.0	0.0	0.0	11.8	31.3	20.0	12.1
Kimi K2.5	40.9	0.0	50.0	10.0	0.0	0.0	20.6	43.8	50.0	24.8
Closed-source										
Gemini-3-Pro + UI-Venus-72B	54.6	4.6	56.3	20.0	0.0	11.1	41.2	75.0	40.0	**36.9**
GPT-5.1 + UI-Venus-72B	54.6	0.0	31.3	0.0	0.0	11.1	26.5	56.3	40.0	26.9
Seed 2.0	18.2	4.6	43.8	10.0	0.0	11.1	17.7	31.3	20.0	18.1
Average	**27.7**	**3.7**	**28.2**	**4.0**	**2.0**	**4.4**	**14.4**	**33.2**	**19.0**	**16.4**

Best total SR
36.9%
strongest agent still fails most tasks

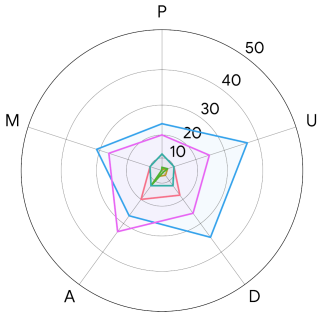
Avg. GSA
2.0%
dynamic state is nearly unsolved

Avg. GUIM
4.4%
visual manipulation collapses

PUDAM Exposes Perception and Memory Bottlenecks



Level 1–2



Level 3–4

Read the drop

- Basic tasks still leave a visible model hierarchy.
- Advanced levels collapse in Perception and Decision.
- Memory remains the cross-model bottleneck.

Same SR can hide different failure causes.

Stability: Small Realistic Variations Break Current Agents

	Environment Settings					Summary		
Model	Original	Chinese	Variation	Dark	Pad	Min	Max	**SPR**
UI-Venus-72B	15	15	20	15	5	5	20	0
UI-Venus-7B	15	20	15	10	5	5	20	**5**
Qwen3-VL-30B-A3B	10	10	15	10	5	5	15	0
Qwen3-VL-8B	10	20	15	10	10	10	20	0
GUI-Owl-7B	15	20	10	15	0	0	20	0
MA3 (GUI-Owl-7B)	10	15	15	20	10	10	20	0
Gemini-3-Pro + UI-Venus-72B	35	35	25	35	30	25	35	**15**
GPT-5.1 + UI-Venus-72B	40	15	35	30	20	15	40	**5**

Stability Pass Rate requires success across all five settings

Most agents have SPR = 0; even the strongest configuration reaches only 15%.

Agentic Reasoning Adds a Deployment Cost

Model	Total	Per-Step
UI-Venus-72B	850.0K	153.2
UI-Venus-7B	447.4K	101.5
Qwen3-VL-30B-A3B	463.2K	138.9
Qwen3-VL-8B	357.7K	132.4
GUI-Owl-7B	373.7K	146.5
MA3 (GUI-Owl-7B)	1640.0K	438.7
Gemini-3-Pro + UI-Venus-72B	259.7K	86.3
GPT-5.1 + UI-Venus-72B	167.5K	54.6

Why token cost?

It is hardware-independent and directly related to API billing and deployment cost.

Framework overhead

MA3 consumes 438.7 output tokens per step, roughly 3.0× GUI-Owl-7B.

Deployment implication

Better reasoning must be balanced with latency, bandwidth, and battery constraints.

What VenusBench-Mobile Changes

1. **Benchmark design:** evaluate mobile GUI agents from user intents, not isolated app functions.
2. **Diagnostic evaluation:** use PUDAM to locate failures in Perception, Understanding, Decision, Action, and Memory.
3. **Empirical finding:** current agents remain far from reliable real-world mobile assistance, especially under dynamic states and environment variations.

VenusBench-Mobile

A challenging, reproducible online benchmark for building more robust general-purpose mobile GUI agents.

Thank you!

Questions?

VenusBench-Mobile: A Challenging and User-Centric Benchmark for Mobile GUI Agents with Capability Diagnostics

Presenter: **Zhuohan Cai**

Contact: **Zhuohan Cai** <caizh24@mails.tsinghua.edu.cn> Yichen Gong <gongyc18@gmail.com>

Tsinghua University Ant Group ICML 2026

Code and data

https://github.com/inclusionAI/UI-Venus/tree/VenusBench-Mobile