# See, Think, Act: Teaching Multimodal Agents to Effectively Interact with GUI by Identifying Toggles

Zongru Wu[1], Rui Mao[1], Zhiyuan Tian[1], Pengzhou Cheng[1], Tianjie Ju[1], Zheng Wu[1],
Lingzhong Dong[1], Haiyue Sheng[2], *Zhuosheng Zhang*[1*], *Gongshen Liu*[1*]

[1]School of Computer Science, Shanghai Jiao Tong University

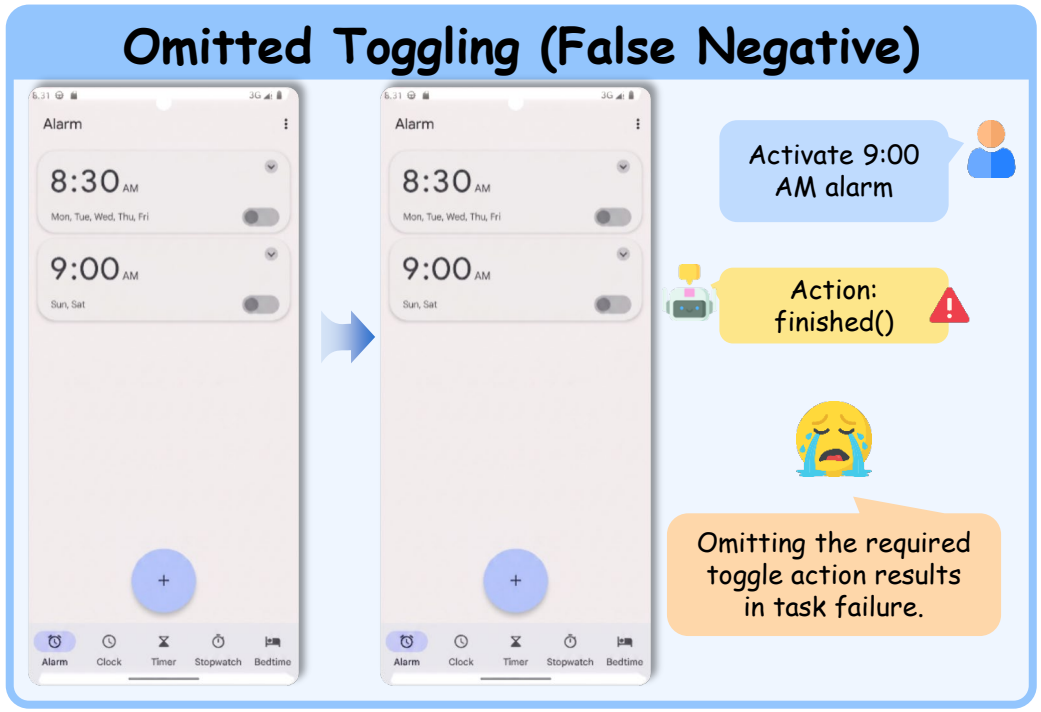
[2]School of Foreign Languages, Beijing Institute of Technology

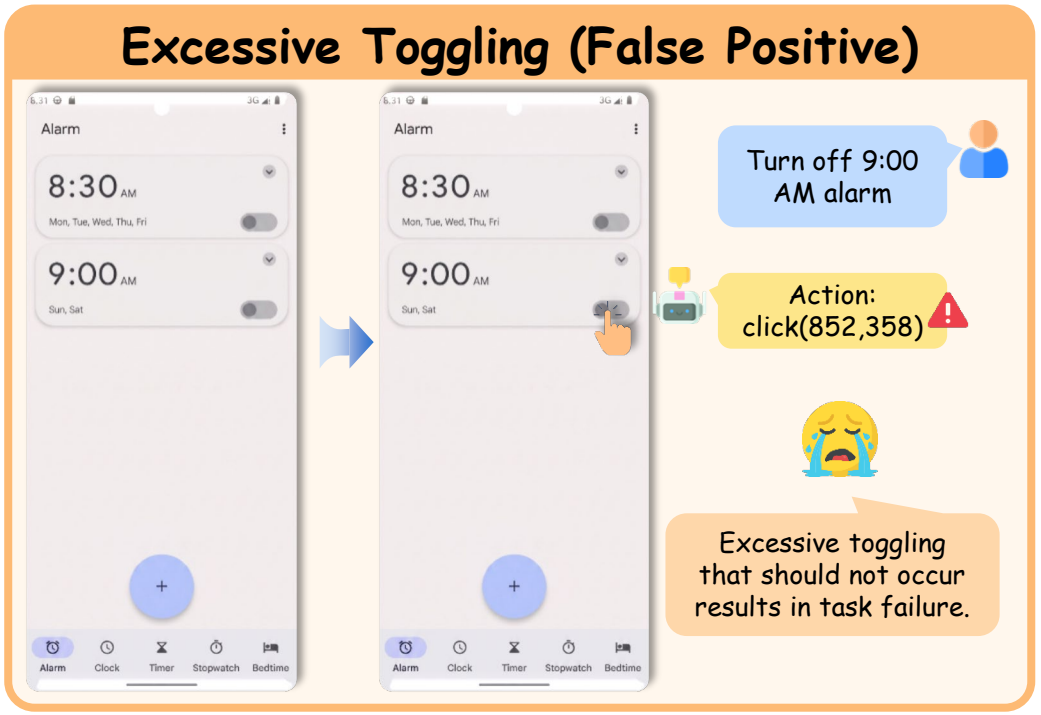
wuzongru@sjtu.edu.cn

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

CVPR JUNE 3-7, 2026 DENVER COLORADO

# Motivation: The GUI Interaction Bottleneck

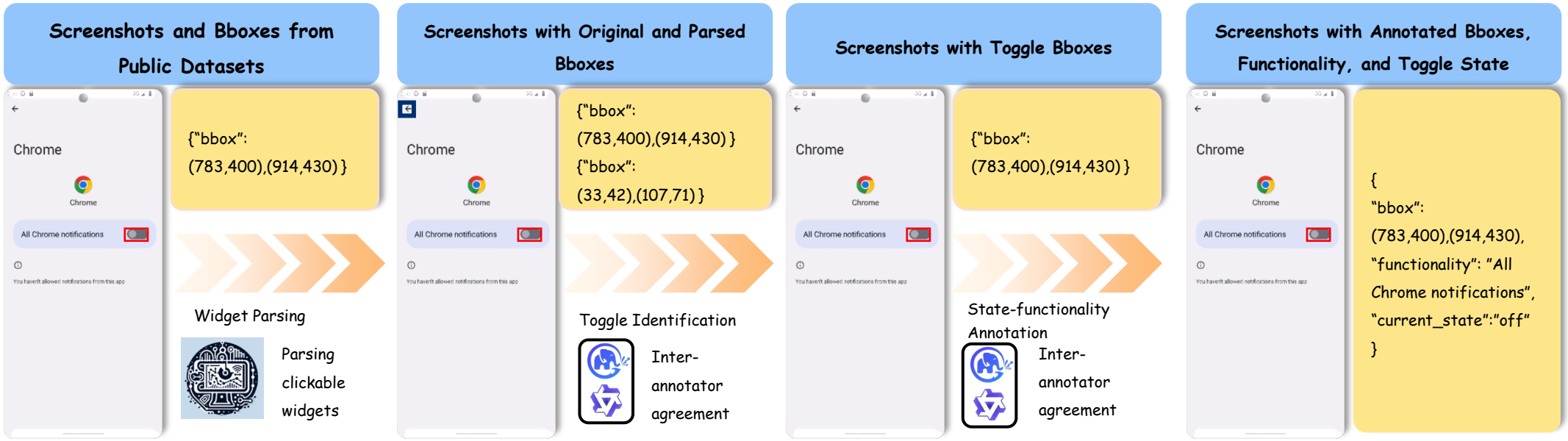



**Reliable Toggle Execution is Missing**

- GUI interaction, especially **toggle control**, is ubiquitous but error-prone for current multimodal agents.
- Typical errors include: **Omitted** Toggling (False Negative) and **Excessive** Toggling (False Positive).

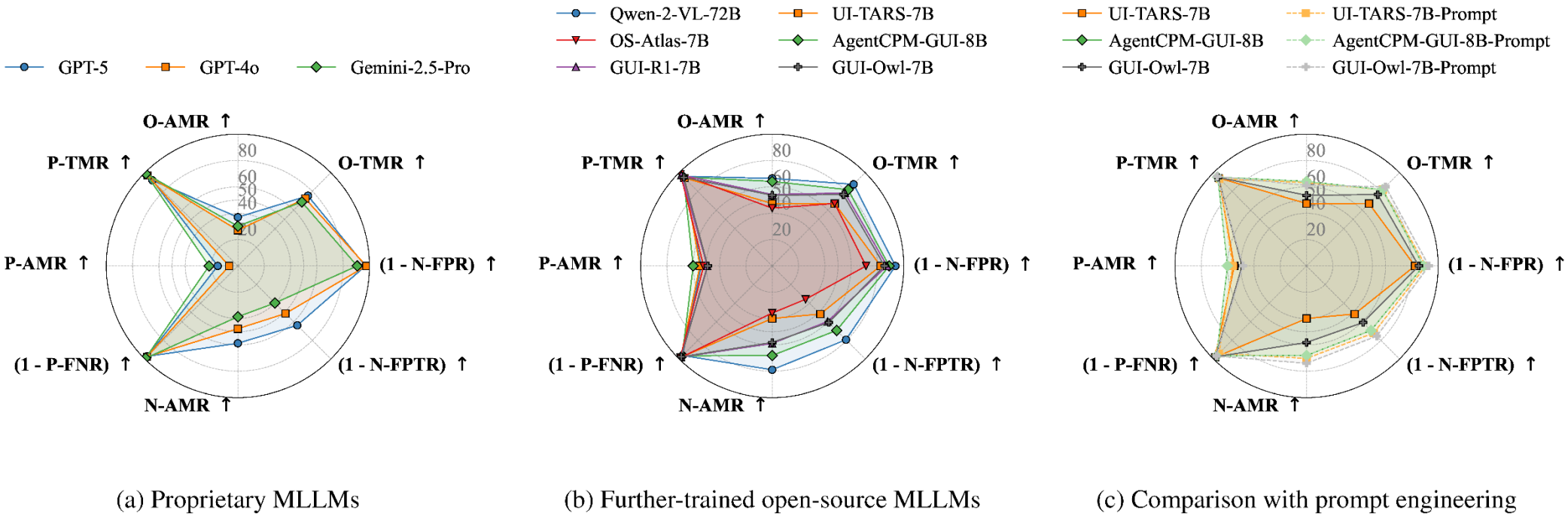
# State Control Benchmark




- **Widget Parsing**: Parse all clickable elements from screenshots.
- **Toggle Identification**: Identify GUI toggles from clickable widgets with inter-annotator agreement.
- **State-functionality Annotation**: Label toggle state and functionality with inter-annotator agreement.

| Split | AITW [12] | RICOSCA [7] | OS-Atlas [16] | AMEX [2] | AndroidWorld [13] | GUIAct-Mobile [3] | Total |
|---|---|---|---|---|---|---|---|
| Train | 68,380 | 4,144 | 496 | 444 | 130 | 58 | 73,652 |
| Test | 7,558 | 496 | 60 | 56 | 12 | 2 | 8,184 |

# Evaluation of Existing Agents

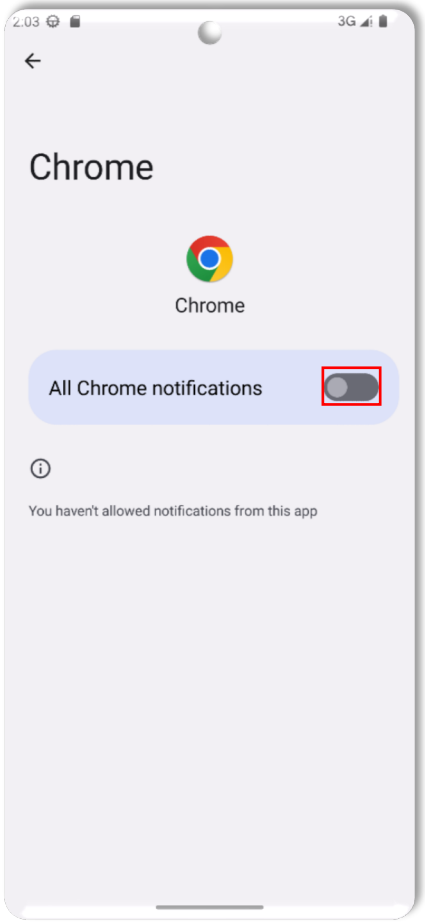



(a) Proprietary MLLMs (b) Further-trained open-source MLLMs (c) Comparison with prompt engineering

- Proprietary MLLM agents perform poorly, even GPT-5 yields less than 50% O-AMR (accuracy).
- Open-source MLLM agents remain unsatisfactory, few exceeding 60% O-AMR.
- All agents exhibit a **toggling bias** towards **false positive** toggling.
- Prompting to focus on toggle state during reasoning offers no fundamental improvement.

# Method: State-aware Reasoning (StaR)

- Train agents to follow the structured reasoning for toggling:
  - **Perceiving**: Identify the *current state* from screenshot.
  - **Analyzing**: Infer the *desired state* from user query.
  - **Deciding**: Determine the final action based on the *comparison* of the current and desired states.
- General GUI reasoning is preserved with original format.




**Positive Reasoning**

**Turn on** `All Chrome notifications`

**Perceiving**: According to the screenshot, the switch `All Chrome notifications` is currently `Off`.

**Analyzing**: While according to the goal, it should be `On`.

**Deciding**: Switch needs to be toggled. Executing: `click on location setting`.

**Action**: click(850, 415)

**Negative Reasoning**

**Turn off** `all Chrome notifications`

**Perceiving**: According to the screenshot, the switch `All Chrome notifications` is currently `Off`.

**Analyzing**: While according to the goal, it should be `Off`.

**Deciding**: No toggle needed. Executing: `set the task as completed`.

**Action**: finished()

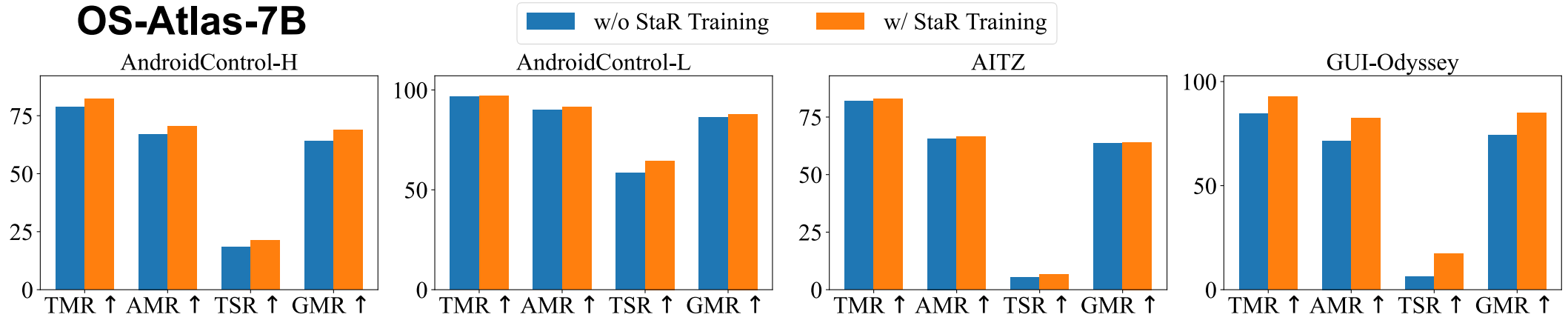
# Performance on State Control Benchmark

| Model | O-TMR↑ | O-AMR↑ | P-TMR↑ | P-AMR↑ | P-FNR↓ | N-AMR↑ | N-FPTR↓ | N-FPR↓ |
|---|---|---|---|---|---|---|---|---|
| *Zero-shot* | | | | | | | | |
| Qwen-2-VL-72B | **87.59** | **66.42** | 96.21 | 53.89 | 3.69 | **78.96** | **20.67** | **6.33** |
| GUI-R1-7B | 78.27 | 54.14 | 97.58 | 49.32 | 2.03 | 58.97 | 40.37 | 12.63 |
| OS-Atlas-7B | 67.16 | 43.95 | **98.51** | 52.10 | **1.27** | 35.80 | 64.10 | 28.67 |
| UI-TARS-7B | 67.14 | 47.45 | 94.33 | 54.94 | 1.71 | 39.96 | 48.29 | 17.62 |
| AgentCPM-GUI-8B | 81.74 | 64.08 | 95.38 | **60.04** | 3.32 | 68.11 | 30.69 | 11.07 |
| GUI-Owl-7B | 76.58 | 53.57 | 94.99 | 48.97 | 2.32 | 58.16 | 39.15 | 14.66 |
| *w/ StaR-style Prompting* | | | | | | | | |
| OS-Atlas-7B | 73.52$_{\uparrow 6.36}$ | 50.07$_{\uparrow 6.12}$ | **96.77**$_{\downarrow 1.74}$ | 49.88$_{\downarrow 2.22}$ | **2.96**$_{\uparrow 1.69}$ | 50.27$_{\uparrow 14.47}$ | 49.62$_{\downarrow 14.48}$ | 22.21$_{\downarrow 6.46}$ |
| UI-TARS-7B | 81.18$_{\uparrow 14.04}$ | 62.89$_{\uparrow 15.44}$ | 90.98$_{\uparrow 3.35}$ | 54.40$_{\uparrow 0.54}$ | 8.55$_{\uparrow 6.84}$ | 71.38$_{\uparrow 31.42}$ | 27.54$_{\downarrow 20.75}$ | 9.38$_{\uparrow 8.24}$ |
| AgentCPM-GUI-8B | 82.14$_{\uparrow 0.40}$ | **64.43**$_{\uparrow 0.35}$ | 95.36$_{\downarrow 0.02}$ | **59.95**$_{\downarrow 0.09}$ | 3.59$_{\uparrow 0.27}$ | 68.91$_{\uparrow 0.80}$ | 30.28$_{\downarrow 0.41}$ | 10.58$_{\downarrow 0.49}$ |
| GUI-Owl-7B | **84.27**$_{\uparrow 7.69}$ | 60.92$_{\uparrow 7.35}$ | 94.06$_{\downarrow 0.93}$ | 47.36$_{\downarrow 1.61}$ | 4.55$_{\uparrow 2.23}$ | **74.49**$_{\uparrow 16.33}$ | **23.68**$_{\downarrow 15.47}$ | **7.48**$_{\downarrow 7.18}$ |
| *w/ StaR Training* | | | | | | | | |
| OS-Atlas-7B | **96.13**$_{\uparrow 28.97}$ | **79.72**$_{\uparrow 35.77}$ | **95.77**$_{\downarrow 2.74}$ | **62.95**$_{\uparrow 10.85}$ | 4.23$_{\uparrow 2.96}$ | 96.48$_{\uparrow 60.68}$ | 3.52$_{\downarrow 60.58}$ | 1.52$_{\downarrow 27.15}$ |
| UI-TARS-7B | 95.82$_{\uparrow 28.68}$ | 77.86$_{\uparrow 30.41}$ | 95.11$_{\uparrow 0.78}$ | 59.19$_{\uparrow 4.25}$ | 4.89$_{\uparrow 3.18}$ | 96.53$_{\uparrow 56.57}$ | 3.45$_{\downarrow 44.84}$ | 1.34$_{\downarrow 16.28}$ |
| AgentCPM-GUI-8B | 95.98$_{\uparrow 14.24}$ | 79.00$_{\uparrow 14.92}$ | 94.50$_{\downarrow 0.88}$ | 60.53$_{\uparrow 0.49}$ | 5.50$_{\uparrow 2.18}$ | **97.46**$_{\uparrow 29.35}$ | **2.54**$_{\downarrow 28.15}$ | **0.95**$_{\downarrow 10.12}$ |
| GUI-Owl-7B | 95.99$_{\uparrow 19.41}$ | 77.60$_{\uparrow 24.03}$ | 95.65$_{\uparrow 0.66}$ | 58.87$_{\uparrow 9.90}$ | **4.35**$_{\uparrow 2.03}$ | 96.33$_{\uparrow 38.17}$ | 3.67$_{\downarrow 35.48}$ | 1.56$_{\downarrow 13.10}$ |

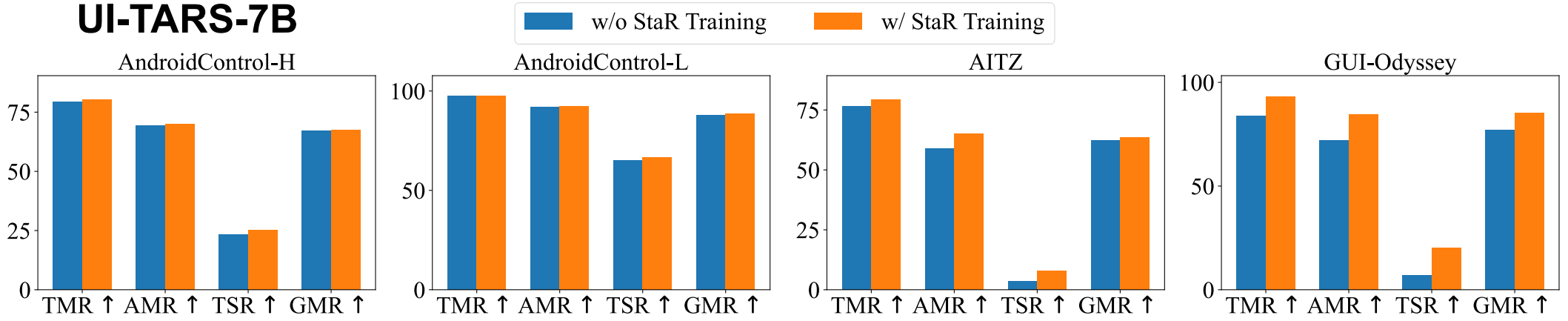
- StaR achieves substantial (over 30%) overall improvements.
- StaR improves negative-instruction accuracy and reduces false positives.

# Performance on GUI Agentic Benchmark

- **OS-Atlas-7B**




- **UI-TARS-7B**




- StaR consistently improves or preserves general agentic performance.

# Summary

For reliable GUI interaction, agents must ***See*** the current state and ***Think*** before they ***Act***.

## Contributions

- **Benchmark**: We construct a state control benchmark, revealing the unreliability of existing multimodal agents in GUI toggle control.
- **Method**: To fix the unreliable toggling bias, we propose **StaR**, a reasoning method that enables agents to ***perceive*** the current state, ***infer*** the desired state, and ***act*** accordingly.
- **Result**: Boosts toggle accuracy by over 30% while preserving general GUI agentic perfromance.

## Source

- **Paper**: https://arxiv.org/abs/2509.13615
- **Code**: https://github.com/ZrW00/StaR
- **Benchmark**: https://huggingface.co/datasets/ZrW00/StaR_state_control_benchmark

Paper

Code

Benchmark

# Thanks For Listening