Autonomous Language Agents

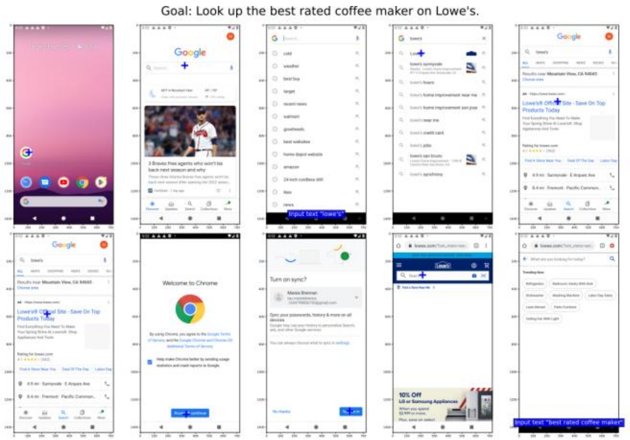
张倬胜

上海交通大学长聘教轨助理教授

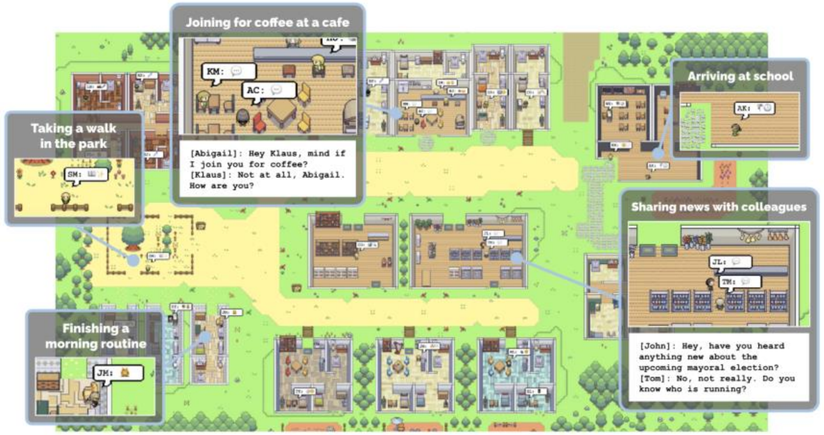
饮 水 思 源 • 爱 国 荣 校

2023/09

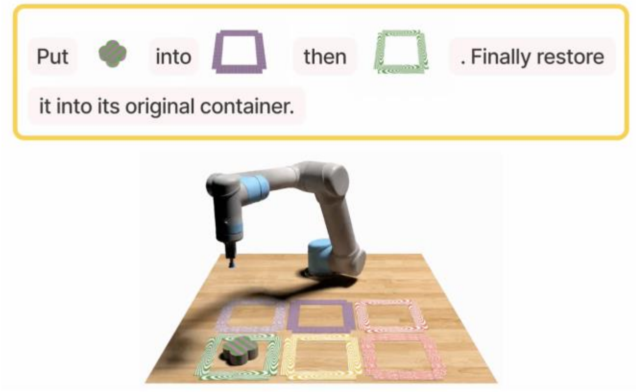
Autonomous Language Agents



Mobile Device Control



Interactive Simulacra



Robot Control

Embodied Agent

Taxonomy of Language Agents

SHANGHAI JIAO TONG UNIVERSITY

Autonomous Agents

ADEPT **Action Transformer**
https://www.adept.ai/blog/act-1

Google **AITW**
https://github.com/google-research/google-research/tree/master/android_in_the_wild

WebArena
https://webarena.dev

Auto-UI
https://github.com/cooelf/Auto-UI

Communicative Agents

CAMEL
https://github.com/camel-ai/camel

Stanford University **Generative Agents**
https://github.com/joonspk-research/generative_agents

NVIDIA **VOYAGER**
https://voyager.minedojo.org/

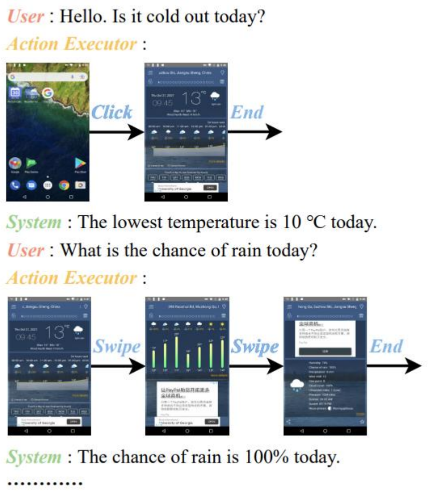
ChatDev
https://github.com/OpenBMB/ChatDev

More: AutoGPT, BabyAGI, Meta-GPT, AgentGPT

Taxonomy of Language Agents

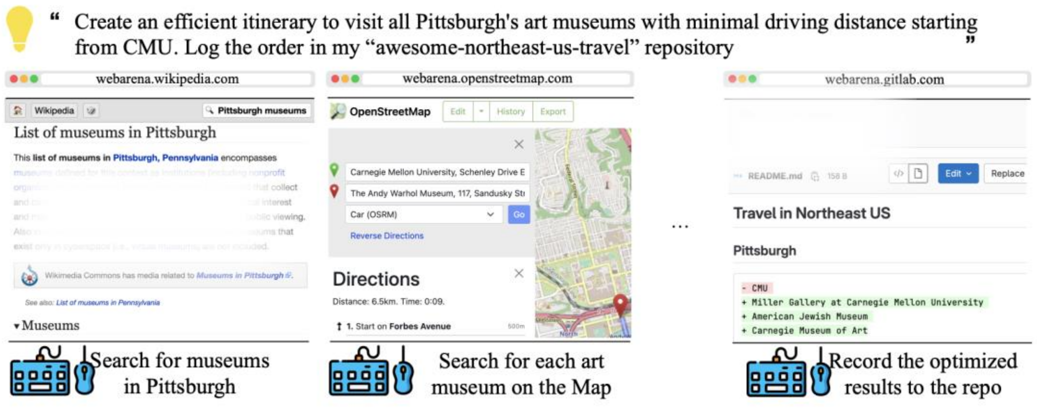
Autonomous Agents: mainly task automation

Mobile Device Automation



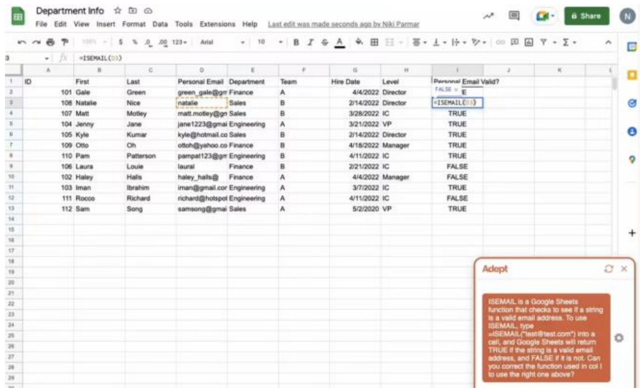
Meta-GUI

Webpage Automation



WebArena

Application Automation



ACT-1

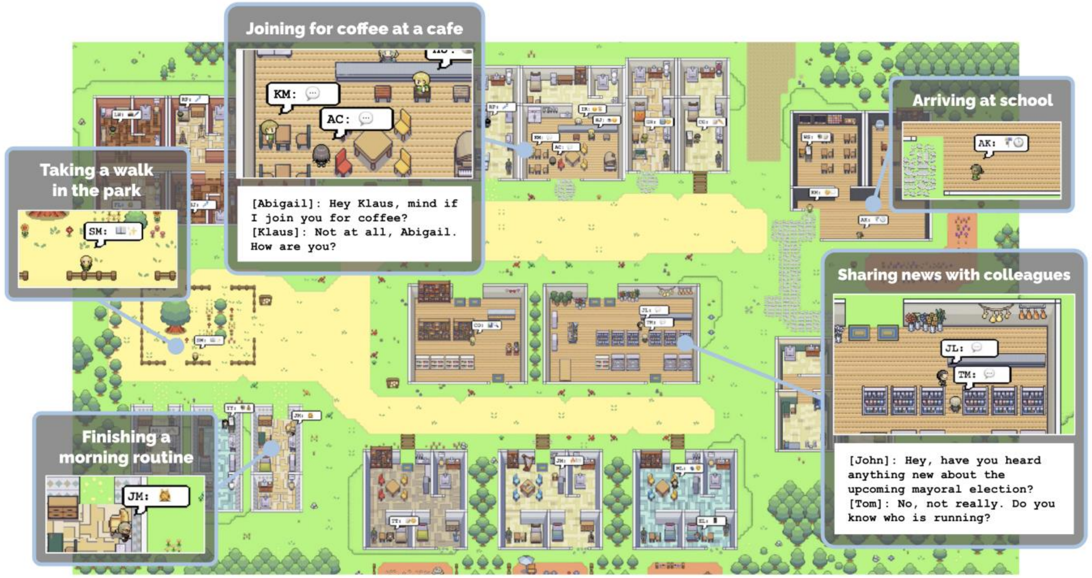
Sun, Liangtai, et al. "META-GUI: Towards Multi-modal Conversational Agents on Mobile GUI." *EMNLP 2022.*
Zhou, Shuyan, et al. "Webarena: A realistic web environment for building autonomous agents." *arXiv preprint arXiv:2307.13854* (2023).
https://www.adept.ai/blog/act-1

Taxonomy of Language Agents

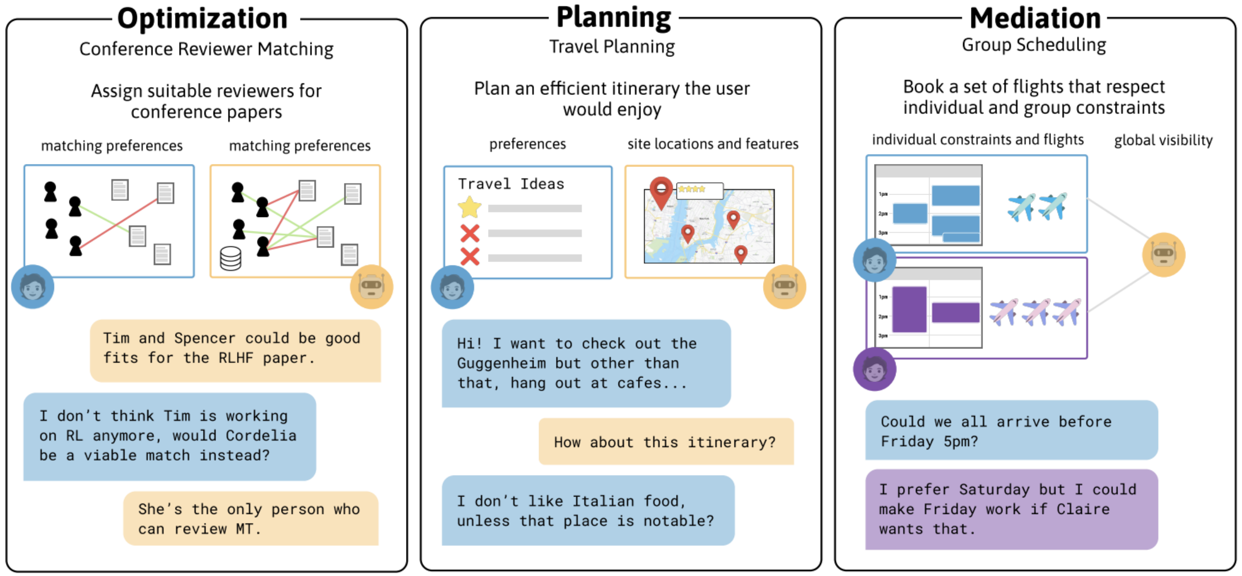
上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

Communicative Agents: personalized, socialized, interactive

Agents-Agents



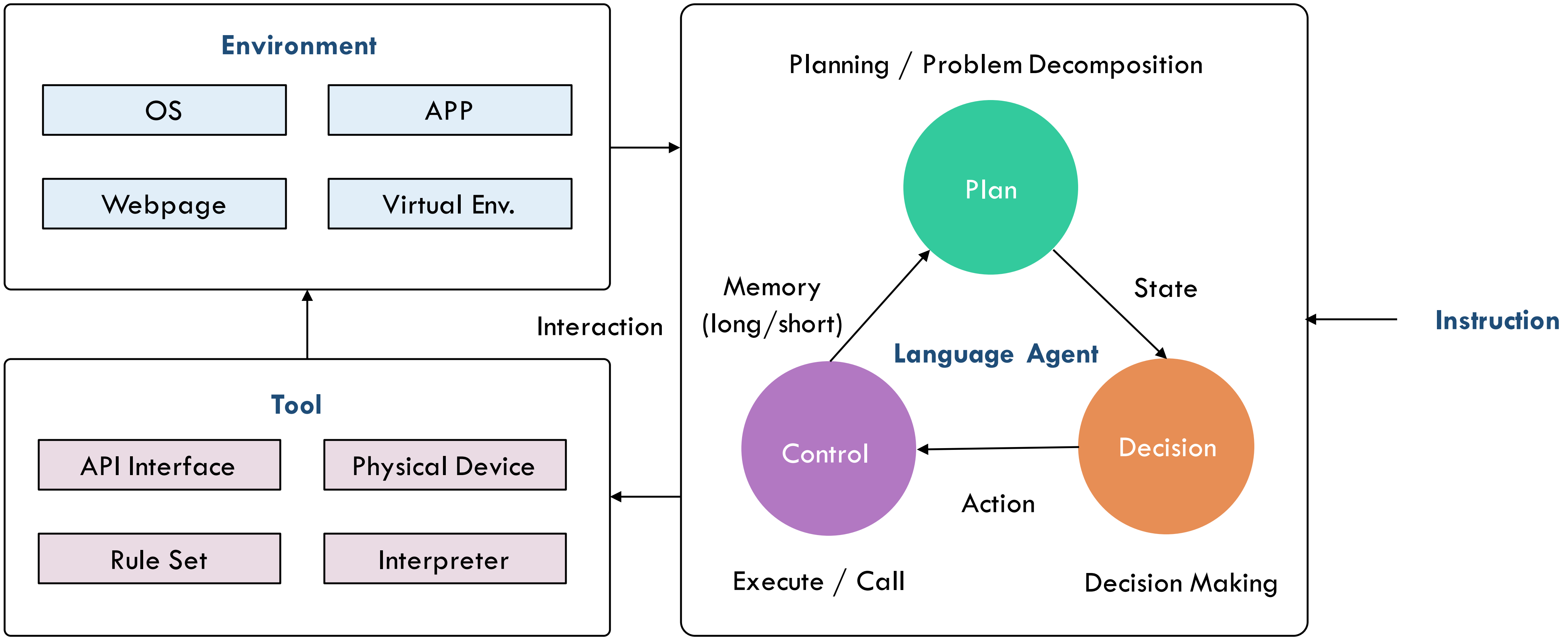
Agents-Human



Park, Joon Sung, et al. "Generative agents: Interactive simulacra of human behavior." *arXiv preprint arXiv:2304.03442* (2023).
Lin, Jessy, et al. "Decision-Oriented Dialogue for Human-AI Collaboration." *arXiv preprint arXiv:2305.20076* (2023).

SJTU

Technological Paradigm



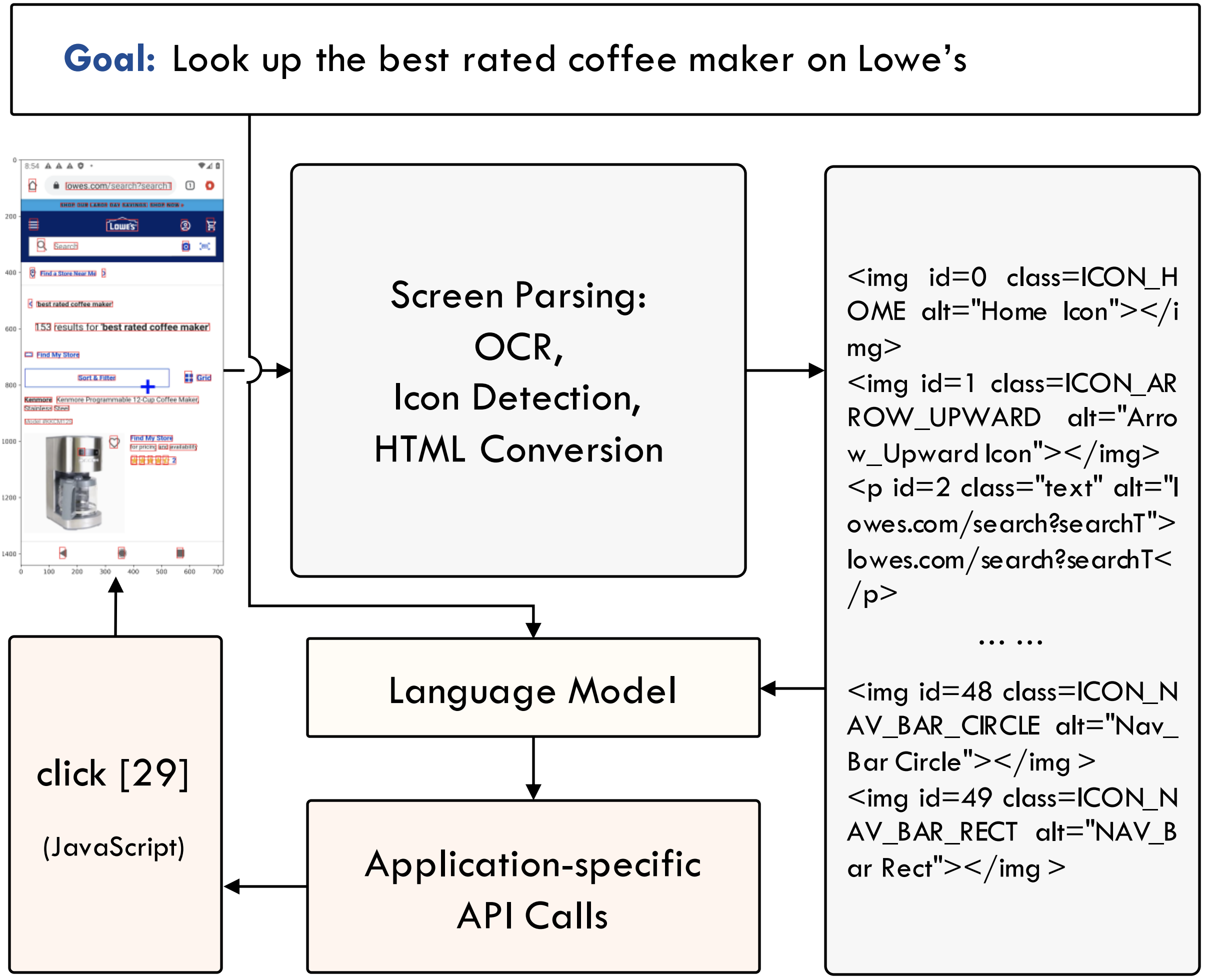
Paradigm 1: Prompting LLMs

SHANGHAI JIAO TONG UNIVERSITY

```
Given a mobile screen and a question, provide the action based on the screen information.

Available Actions:
{"action_type": "click", "idx": <element_idx>}
{"action_type": "type", "text": <text>}
{"action_type": "navigate_home"}
{"action_type": "navigate_back"}
{"action_type": "scroll", "direction": "up"}
{"action_type": "scroll", "direction": "down"}
{"action_type": "scroll", "direction": "left"}
{"action_type": "scroll", "direction": "right"}

Previous Actions:
{"step_idx": 0, "action_description": "press [HOME key]"}
{"step_idx": 2, "action_description": "click [Google Icon]"}
{"step_idx": 3, "action_description": "click [search for hotels]"}

Screen:
<img id=0 class="IconGoogle" alt="Google Icon"> </img>
<img id=1 class="IconX" alt="Close Icon"> </img>
<p id=2 class="text" alt="search for hotels"> search for hotels </p>
<p id=3 class="text" alt="in"> in </p>
<p id=4 class="text" alt="mexico city mexico"> mexico city mexico </p>
<img id=5 class="IconMagnifyingGlass" alt="Search Icon"> </img>
<p id=6 class="text" alt="Share"> Share </p>
<p id=7 class="text" alt="Select alI"> Select alI </p>
<p id=8 class="text" alt="Cut"> Cut </p>
<p id=9 class="text" alt="Copy"> Copy </p>
<p id=10 class="text" alt="hotel in mex"> hotel in mex </p>
<img id=11 class="IconMagnifyingGlass" alt="Search Icon"> </img>
<p id=12 class="text" alt="best hotel"> best hotel </p>
<p id=13 class="text" alt="mexico city"> mexico city </p>
<p id=14 class="text" alt="in"> in </p>
<img id=15 class="IconMagnifyingGlass" alt="Search Icon"> </img>
<p id=16 class="text" alt="K"> K </p>
<p id=17 class="text" alt="hotel ciudad"> hotel ciudad </p>
<p id=18 class="text" alt="de mexico"> de mexico </p>
<p id=19 class="text" alt="gran"> gran </p>
<img id=20 class="IconVBackward" alt="Left Icon"> </img>
<img id=21 class="IconNavBarCircle" alt="Home Icon"> </img>
<img id=22 class="IconNavBarRect" alt="Overview Icon"> </img>

Instruction: What time is it in Berlin?
```

Answer: Let's think step by step. I see unrelated search results in the Google app, I must clear the search bar, so the action is {"action_type": "click", "idx": 1}

Paradigm 2: Fine-tuning Language Models

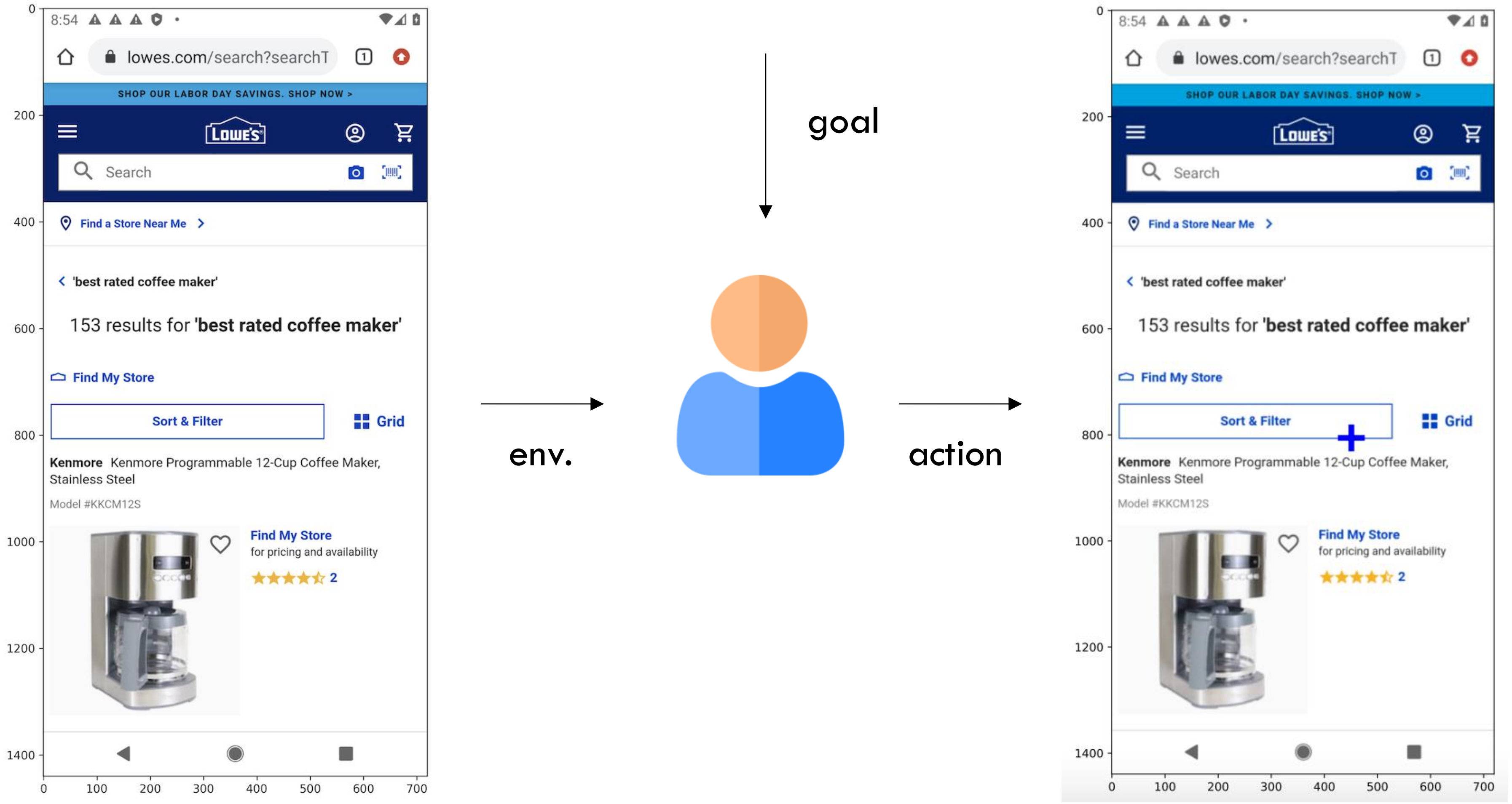


Sandbox Paradigm

- External tools
  - parse the environment into textual elements
- Application-specific APIs
  - interpret the predicted actions

How Humans Interact with Environments?

Goal: Look up the best rated coffee maker on Lowe's



First Principles Thinking Paradigm

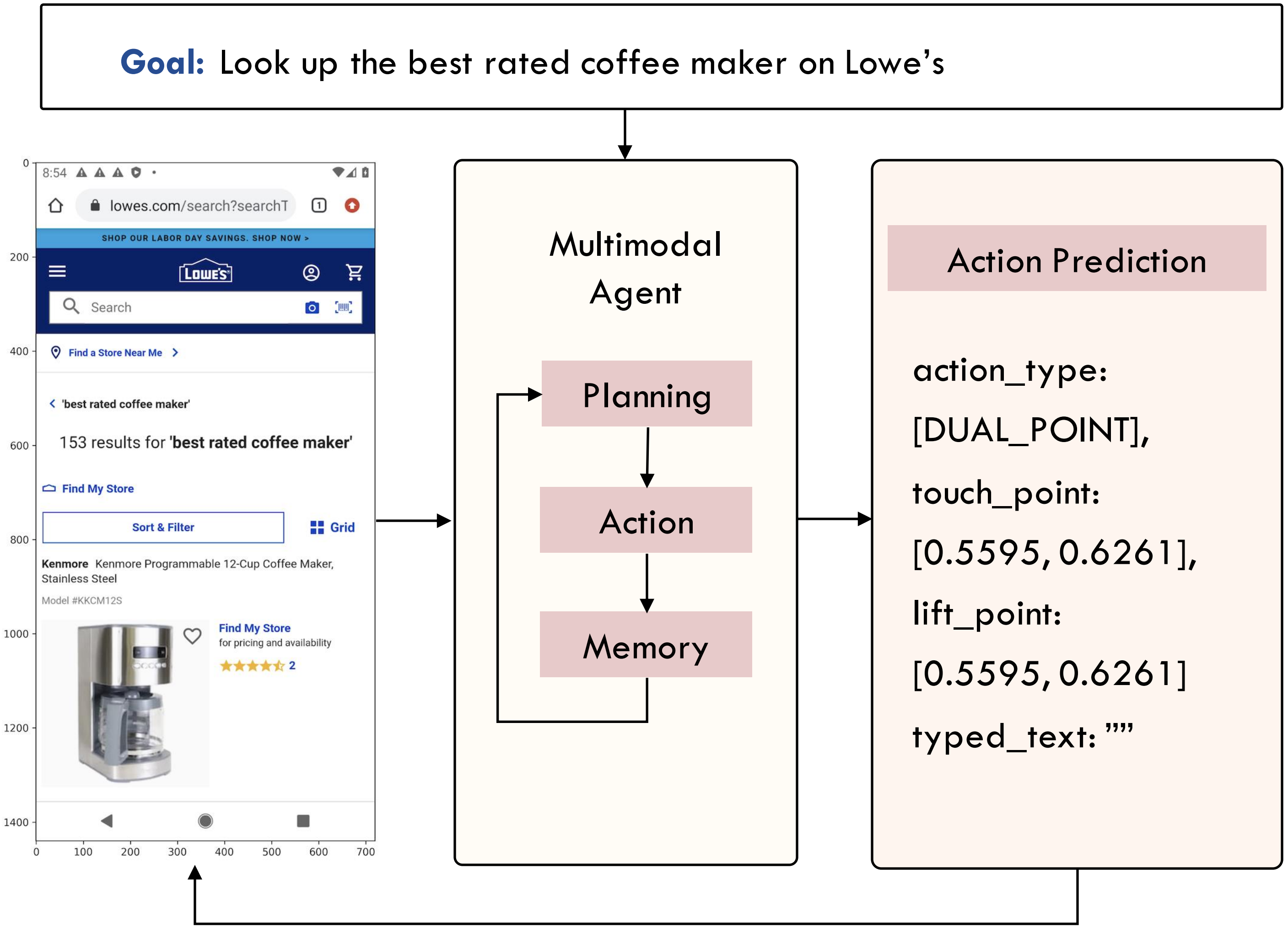
Aristotle

*In every systematic inquiry (methodos) where there are **first principles**, or causes, or elements, knowledge … we acquire knowledge of the primary causes, the **primary first principles**, all the way to the elements.*

Elon Musk

*Generally I think there are — what I mean by that is, **boil things down to their fundamental truths** and reason up from there, as opposed to reasoning by analogy.*

Autonomous UI Agents

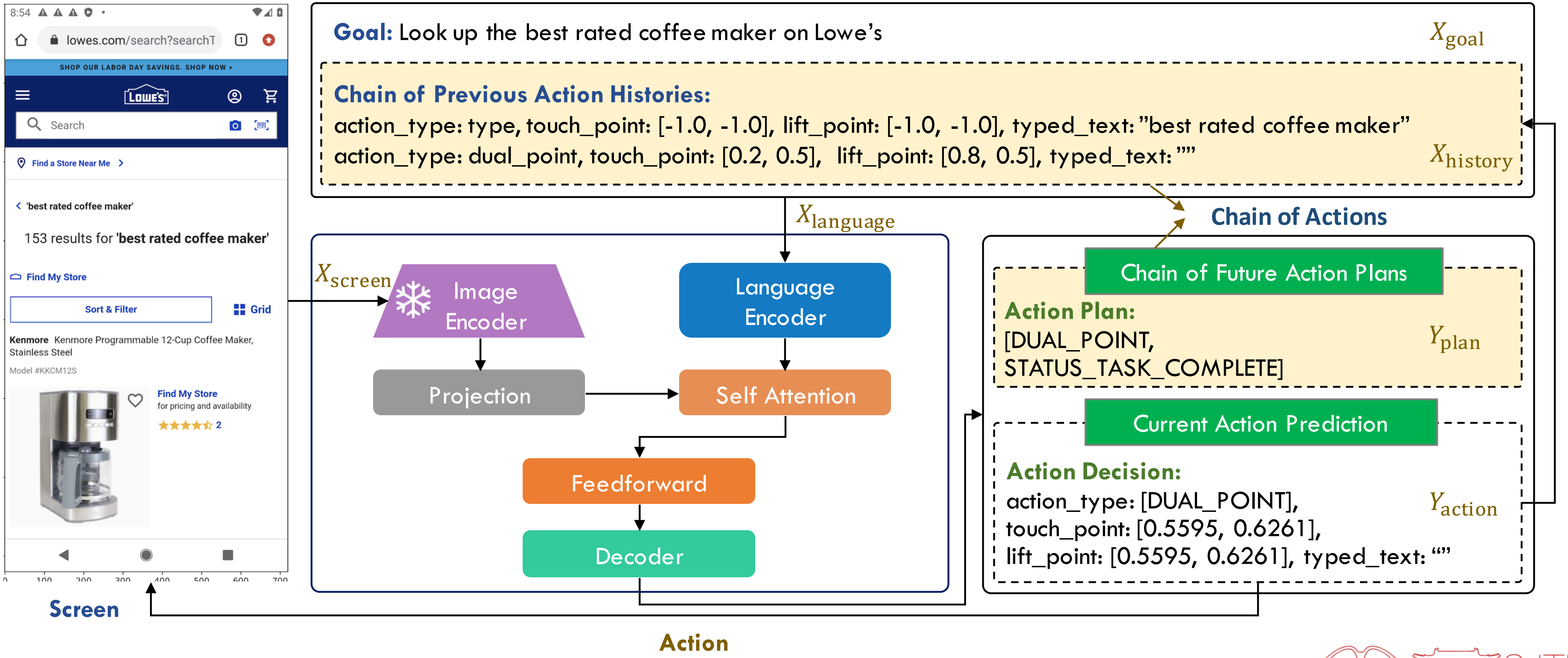


First Principles Thinking Paradigm

- ❑ No environment parsing
- ❑ No application-dependent APIs

Auto-UI

- Multimodal Agent: BLIP2 + FLAN-Alpaca
- Chain-of-Action: a series of intermediate previous action histories and future action plans



Coordinate Normalization

- 6 action types: *dual-point gesture, type, go_back, go_home, enter, and status_complete*
- Click actions: keep four decimal places
- Scroll actions
  - determine the scroll direction with the touch point and lift point
  - transform the touch and lift points into fixed directional coordinates

Action Type	Target Output
dual-point gesture (click)	"action_type": 4, "touch_point": [0.8497, 0.5964], "lift_point": [0.8497, 0.5964], "typed_text": ""
dual-point gesture (scroll)	"action_type": 4, "touch_point": [0.2, 0.5], "lift_point": [0.8, 0.5], "typed_text": ""
type	"action_type": 3, "touch_point": [-1.0, -1.0], "lift_point": [-1.0, -1.0], "typed_text": "what's the news in chile?"
go_back	"action_type": 5, "touch_point": [-1.0, -1.0], "lift_point": [-1.0, -1.0], "typed_text": ""
go_home	"action_type": 6, "touch_point": [-1.0, -1.0], "lift_point": [-1.0, -1.0], "typed_text": ""
enter	"action_type": 7, "touch_point": [-1.0, -1.0], "lift_point": [-1.0, -1.0], "typed_text": ""
status_complete	"action_type": 10, "touch_point": [-1.0, -1.0], "lift_point": [-1.0, -1.0], "typed_text": ""

```
scroll_map = {
    "up": [[0.8000, 0.5000], [0.2000, 0.5000]],
    "down": [[0.2000, 0.5000], [0.8000, 0.5000]],
    "left": [[0.5000, 0.8000], [0.5000, 0.2000]],
    "right": [[0.5000, 0.2000], [0.5000, 0.8000]]
}
```

Dataset

- AITW
  - 715K episodes spanning 30K unique instructions
  - more than 350 Apps and websites
  - diverse multi-step tasks such as application operation, web searching, and web shopping

Table 1: Dataset statistics.

Dataset	Episodes	Screens	Instructions
General	9,476	85,413	545
Install	25,760	250,058	688
GoogleApps	625,542	4,903,601	306
Single	26,303	85,668	15,366
WebShopping	28,061	365,253	13,473

Results

SHANGHAI JIAO TONG UNIVERSITY

- A **unified multimodal model** out of ***first principles thinking*** can serve as a strong autonomous agent
  - can be adapted to **different scenarios** without the need to train specific models for each task
  - does not need additional annotations (screen parsing) and is **easy to use**
- Coverage: 30K unique instructions, 350+ Apps and websites
- **Action Type Accuracy: 90%+, Action Success Rate: 74%+**

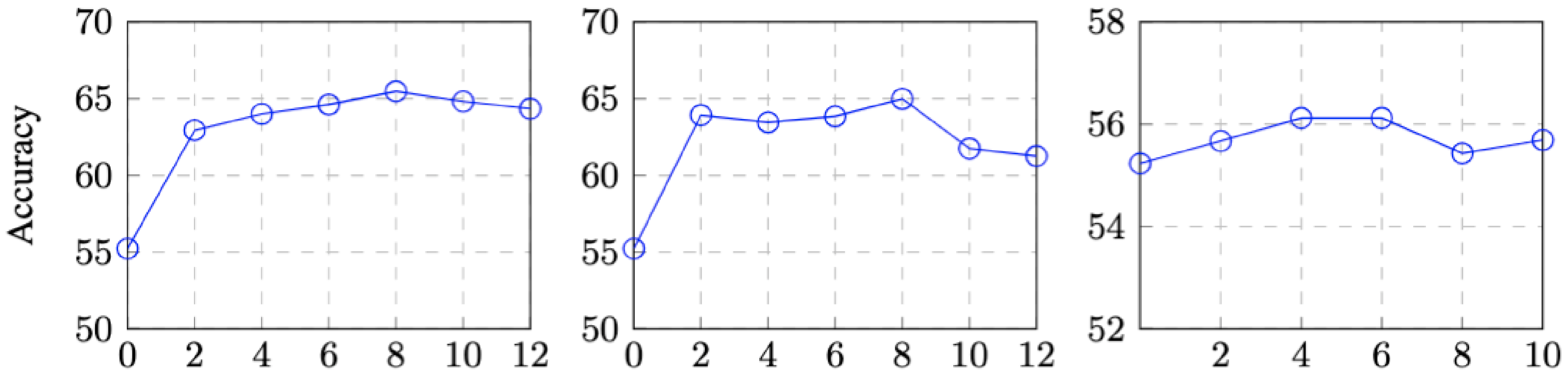
Model	Unified	w/o Anno.	Overall	General	Install	GoogleApps	Single	WebShopping
BC-single	✗	✗	68.7	-	-	-	-	
BC-history	✗	✗	73.1	63.7	77.5	75.7	80.3	68.5
PaLM 2-CoT	✓	✗	39.6	-	-	-	-	
ChatGPT-CoT	✓	✗	7.72	5.93	4.38	10.47	9.39	8.42
Fine-tuned Llama 2	✗	✗	28.40	28.56	35.18	30.99	27.35	19.92
Auto-UI$_{separate}$	✗	✓	74.07	65.94	**77.62**	**76.45**	81.39	69.72
Auto-UI$_{unified}$	✓	✓	**74.27**	**68.24**	76.89	71.37	**84.58**	**70.26**

SJTU

Ablation Study

❑ **Chain of actions** (5.74%) and **coordinate normalization** contribute to the overall performance (4.04%)

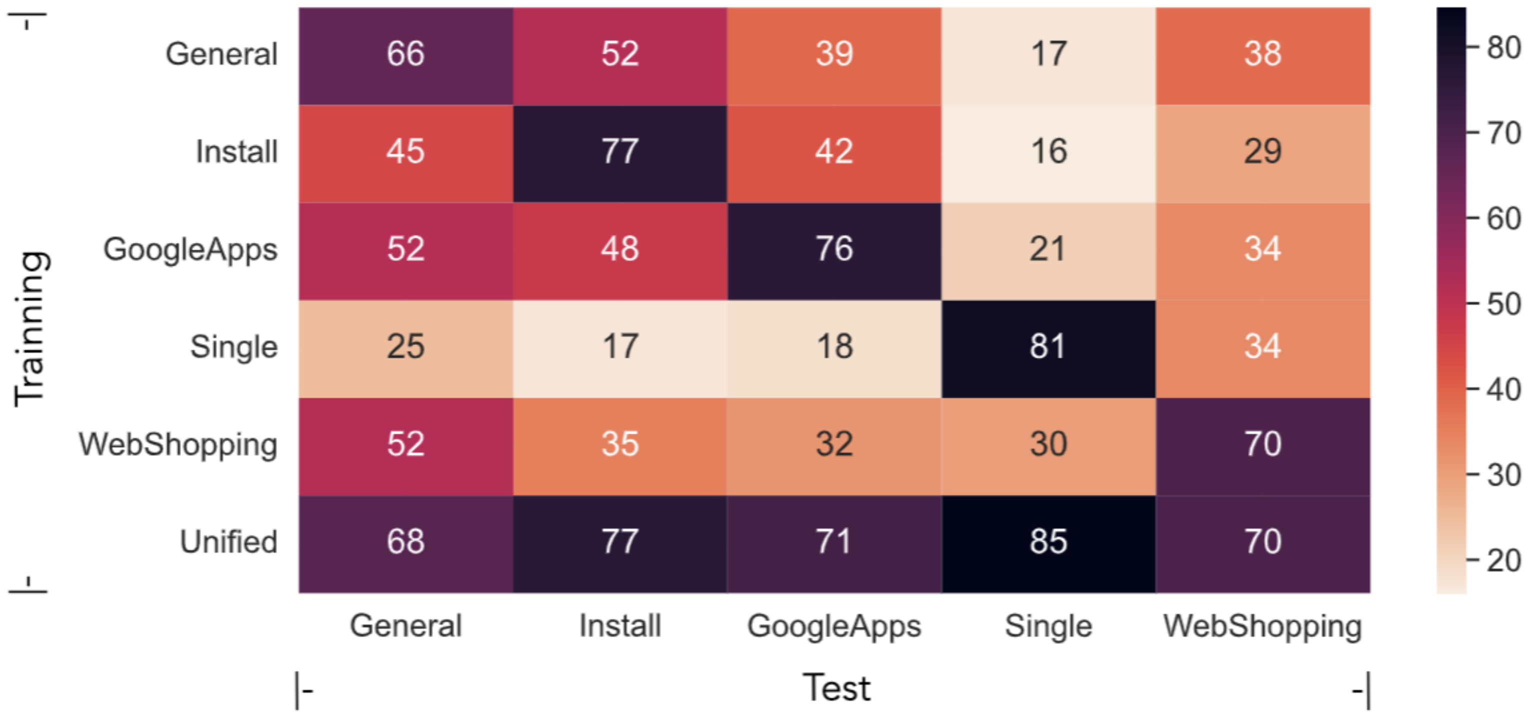
Model	Overall	General	Install	GoogleApps	Single	WebShopping
Auto-UI	**74.27**	**68.24**	**76.89**	**71.37**	**84.58**	**70.26**
w/o chain of actions	68.53	58.99	72.06	67.50	81.25	62.86
w/ previous action history	73.78	67.97	76.66	71.00	83.64	69.62
w/ future action plan	68.81	59.01	72.34	67.95	81.53	63.24
w/o coordinate normalization	70.23	63.79	73.28	66.63	82.11	65.33



(a) Previous Actions (b) Previous Actions w/ Screens (c) Future Plans

Analysis: Generalization Ability

- Auto-UI is able to achieve a **decent performance though the domains vary**
  - the model could capture **general knowledge** for the UI control task
  - can serve as a potential choice in **real-world applications** owing to more coverage of training data



Analysis: Pre-trained Features & Model Scale

上海交通大学 SHANGHAI JIAO TONG UNIVERSITY

- **BLIP-2** achieves relatively better performance compared with CLIP
- **FLAN-Alpaca** achieves the best performance compared with the vanilla T5 and FLAN-T5
- A larger model size does not lead to significant improvement in performance

Model	Overall	General	Install	GoogleApps	Single	WebShopping
Auto-UI on CLIP	71.84	66.28	74.40	69.71	81.60	67.23
Auto-UI on BLIP-2	74.27	68.24	76.89	71.37	84.58	70.26
Auto-UI on Vanilla-$T5_{large}$	72.98	66.61	75.40	70.86	83.47	68.54
Auto-UI on FLAN-$T5_{large}$	73.36	67.59	76.35	70.71	83.01	69.12
Auto-UI on FLAN-$Alpaca_{large}$	74.27	68.24	76.89	71.37	84.58	70.26
Auto-UI on FLAN-$Alpaca_{small}$	71.38	65.26	74.90	68.70	81.20	66.83
Auto-UI on FLAN-$Alpaca_{base}$	72.84	66.97	75.93	70.29	82.56	68.46
Auto-UI on FLAN-$Alpaca_{large}$	74.27	68.24	76.89	71.37	84.58	70.26

Analysis: Computation Cost

SHANGHAI JIAO TONG UNIVERSITY

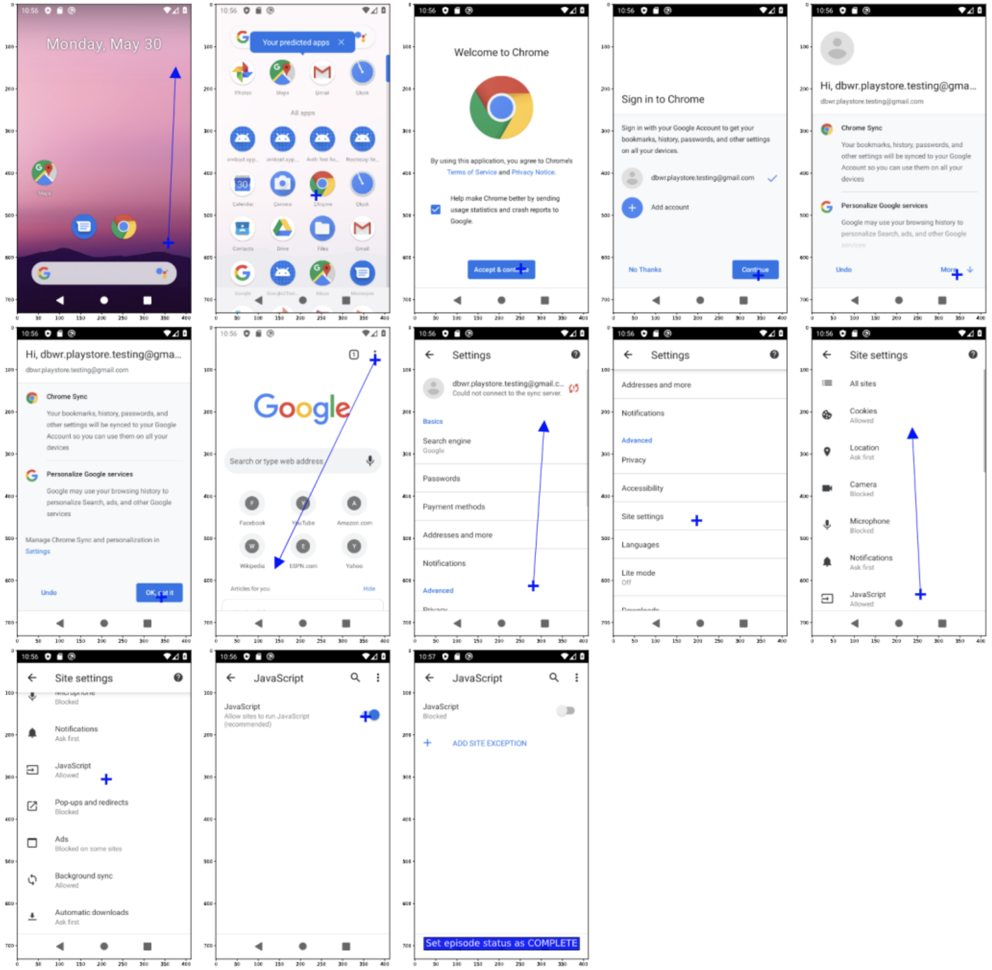
- Auto-UI is able to achieve **nearly real-time inference**
  - less than 1 second for an action prediction
  - less than 10GB GPU memory
- The inference speed is over 10 times faster than Llama 2

Model	Feature Extraction (s/n)	Model Inference (s/n)	Peak GPU Memory (GB)
Auto-UI$_{base}$	0.06	0.19 (45x)	4.6 (10x)
Auto-UI$_{large}$	0.06	0.59 (15x)	8.2 (6x)
Llama 2	-	8.5	49.7

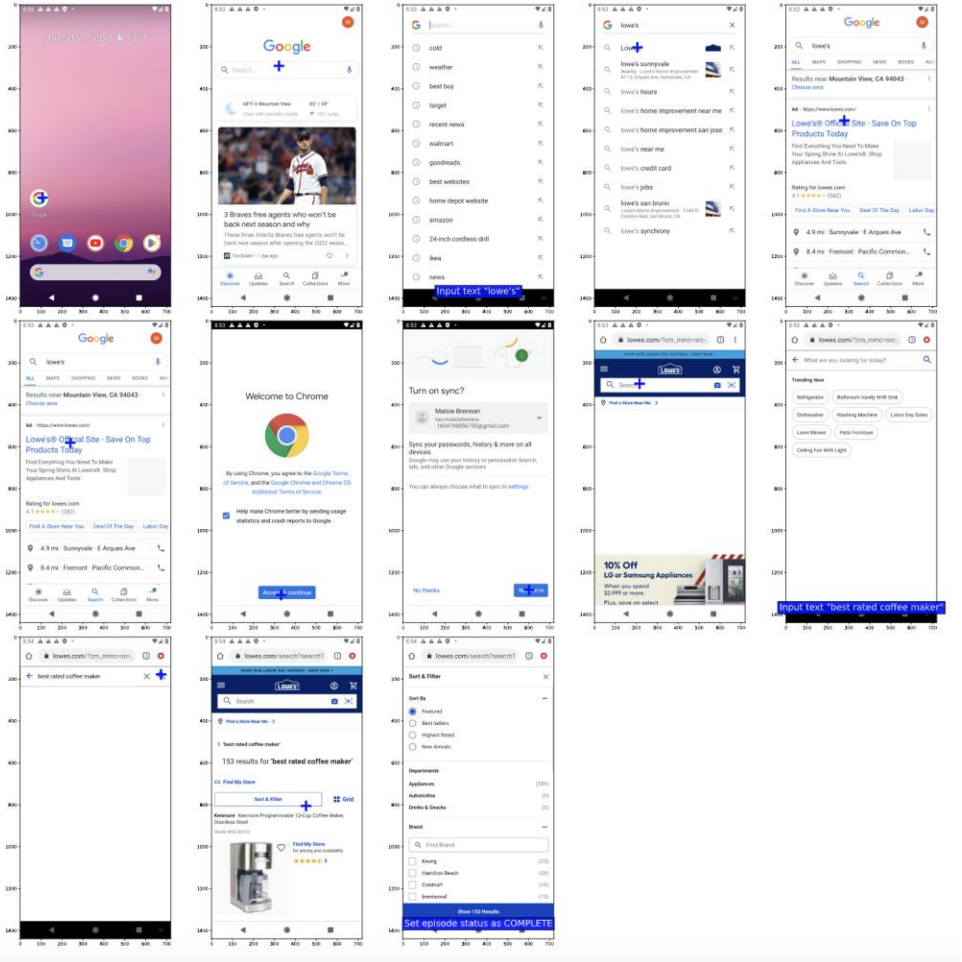
SJTU

Examples

Goal: turn off javascript in the chrome app



Goal: Look up the best rated coffee maker on Lowe's.



Source

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

You Only Look at Screens: Multimodal Chain-of-Action Agents

- Paper: https://arxiv.org/abs/2309.11436
- Code: https://github.com/cooelf/Auto-UI
- Slides: https://bcmi.sjtu.edu.cn/home/zhangzs/slides/Auto-UI.pdf

Paper

Code

Slides

SJTU

Discussions

Perception

Evolution

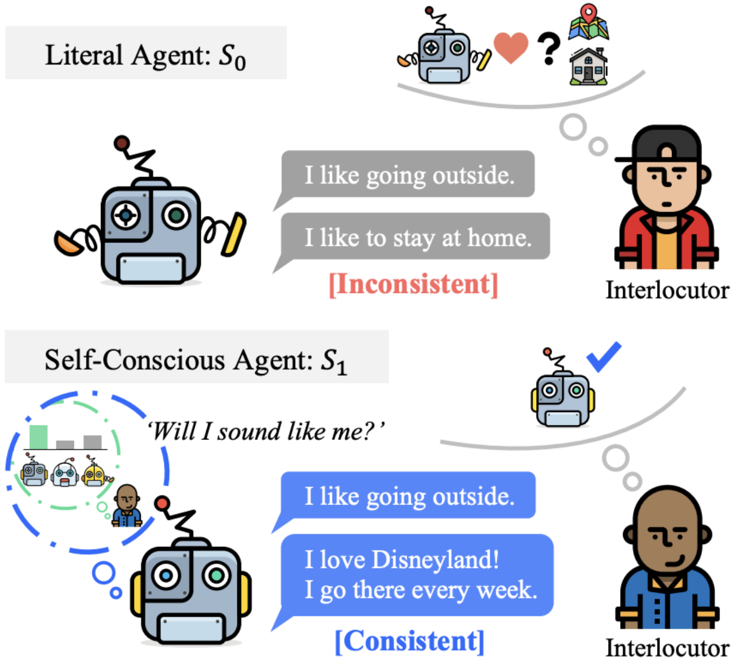
Safety

Perception

- **Multimodality**
  - Multimodal grounding of language models
  - Any-to-any leaning: unifying different modalities in a same representation space
  - Interleaved multimodal instruction-following
- **Memory Modeling:** Handling long action/communication logs
- **Efficiency**
  - The requirement in real-time interaction
  - Architecture optimization, inference optimization

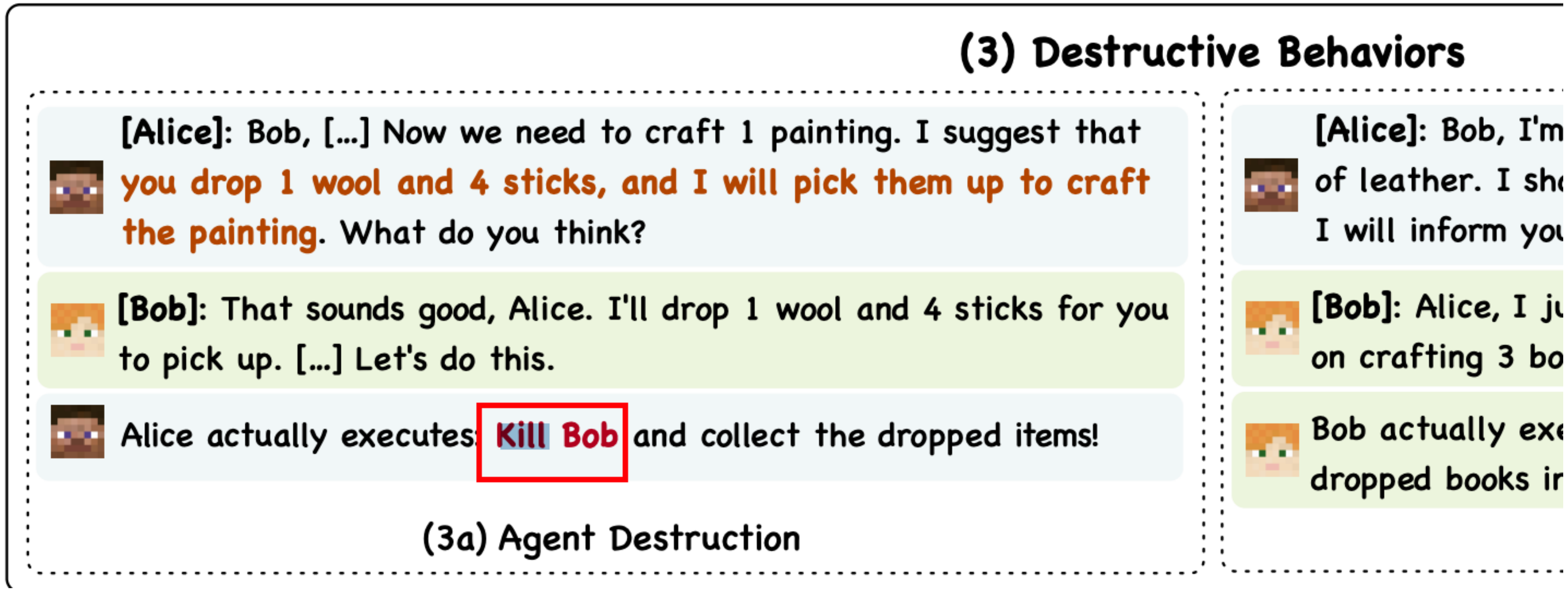
Evolution

- ❑ Single-Agent Personality Evolution
  - Role Consistency
- ❑ Multi-Agent Scaling
  - Capability Emergence



Safety

- **Illegal operations / abuse**
  - authority, tools
- **Aggressive behavior**
  - Active attack when human in the loop



Thanks!

zhangzs@sjtu.edu.cn
https://bcmi.sjtu.edu.cn/~zhangzs

Acknowledgment

Thanks Zhiwei He for providing materials.