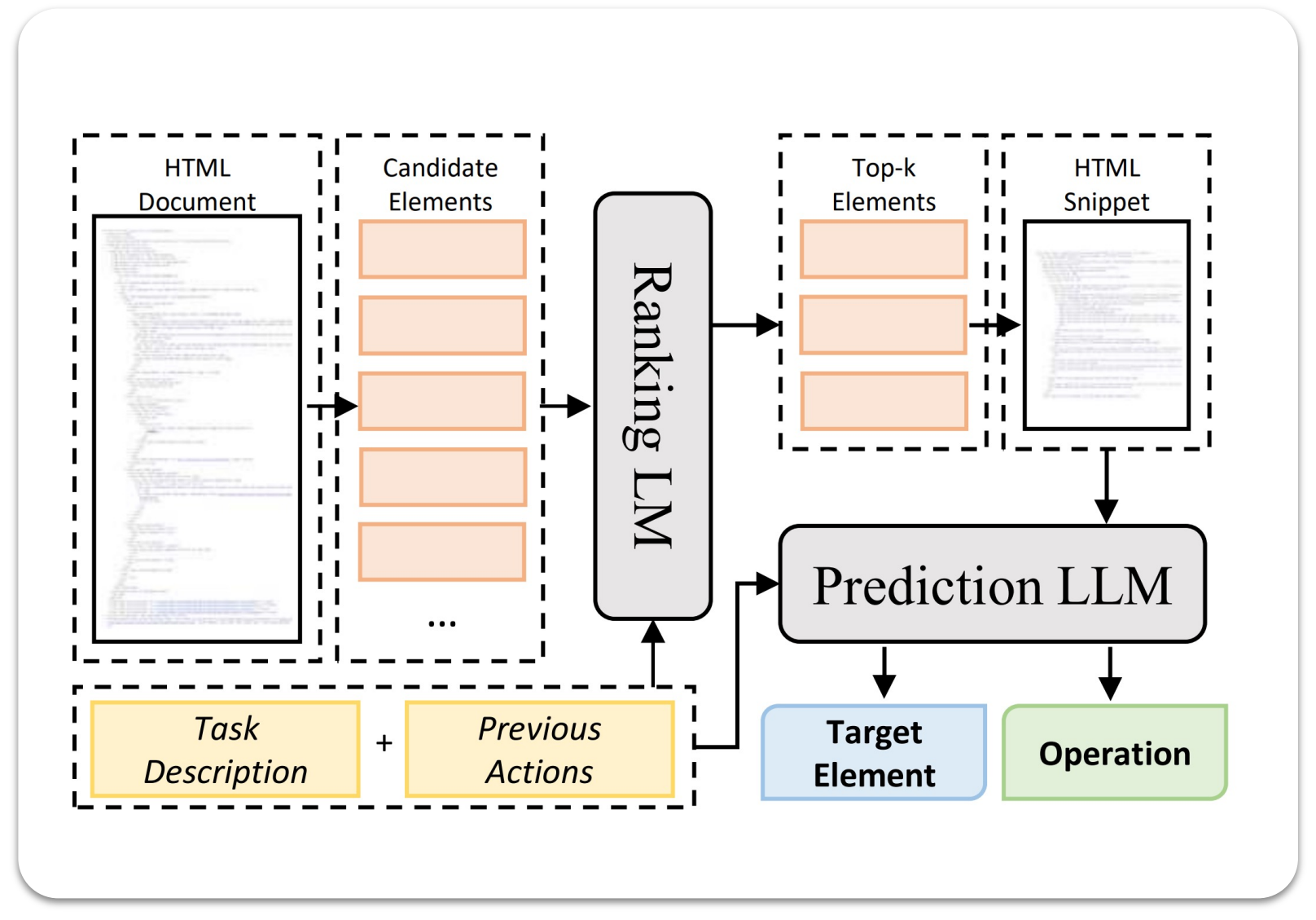
Navigating the Digital World as Humans Do: Universal Visual Grounding For GUI Agents

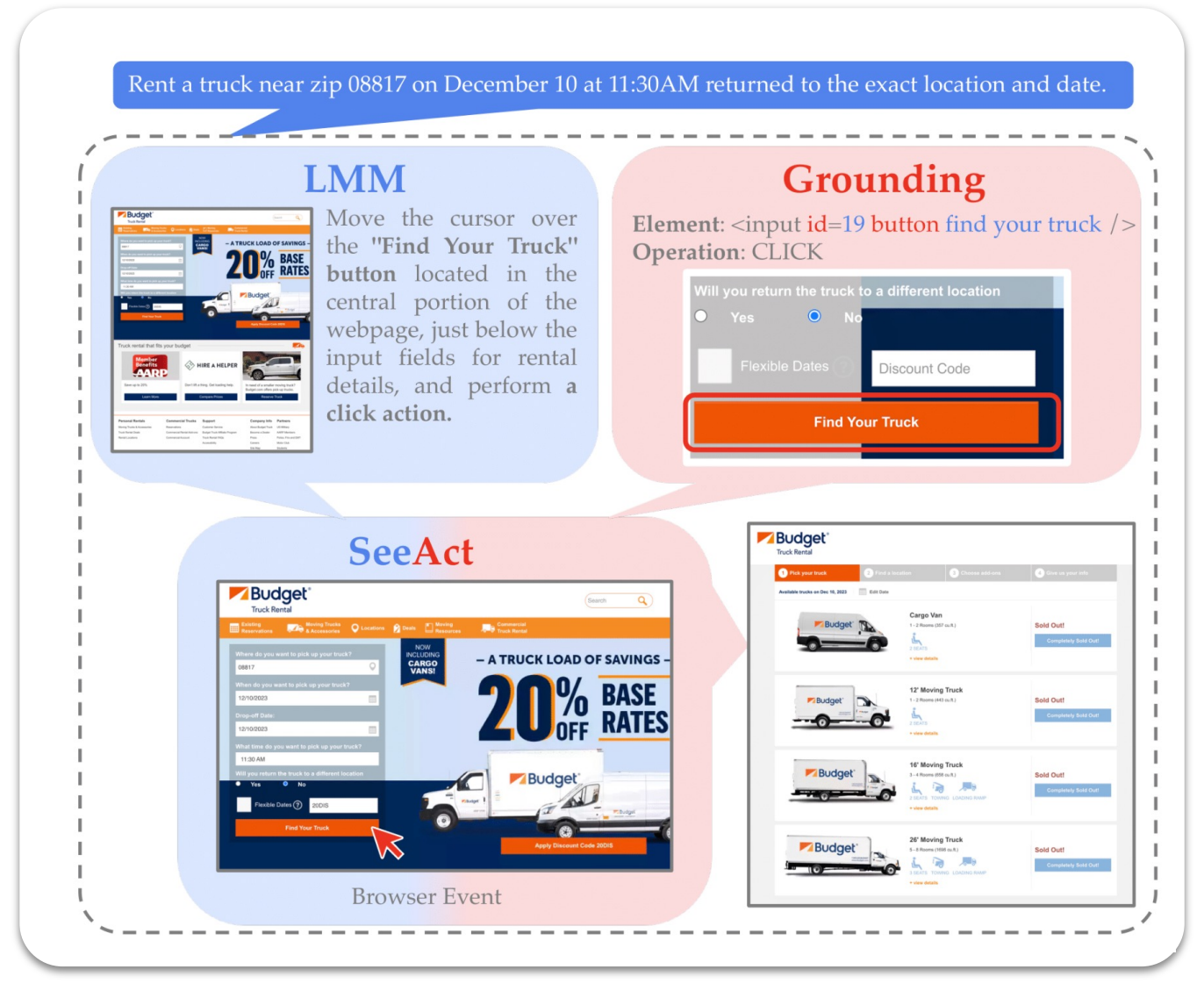
Boyu Gou, Ruohan Wang, Boyuan Zheng, Yanan Xie, Cheng Chang,
Yiheng Shu, Huan Sun, Yu Su

The Ohio State University · NLP · Orby · ICLR

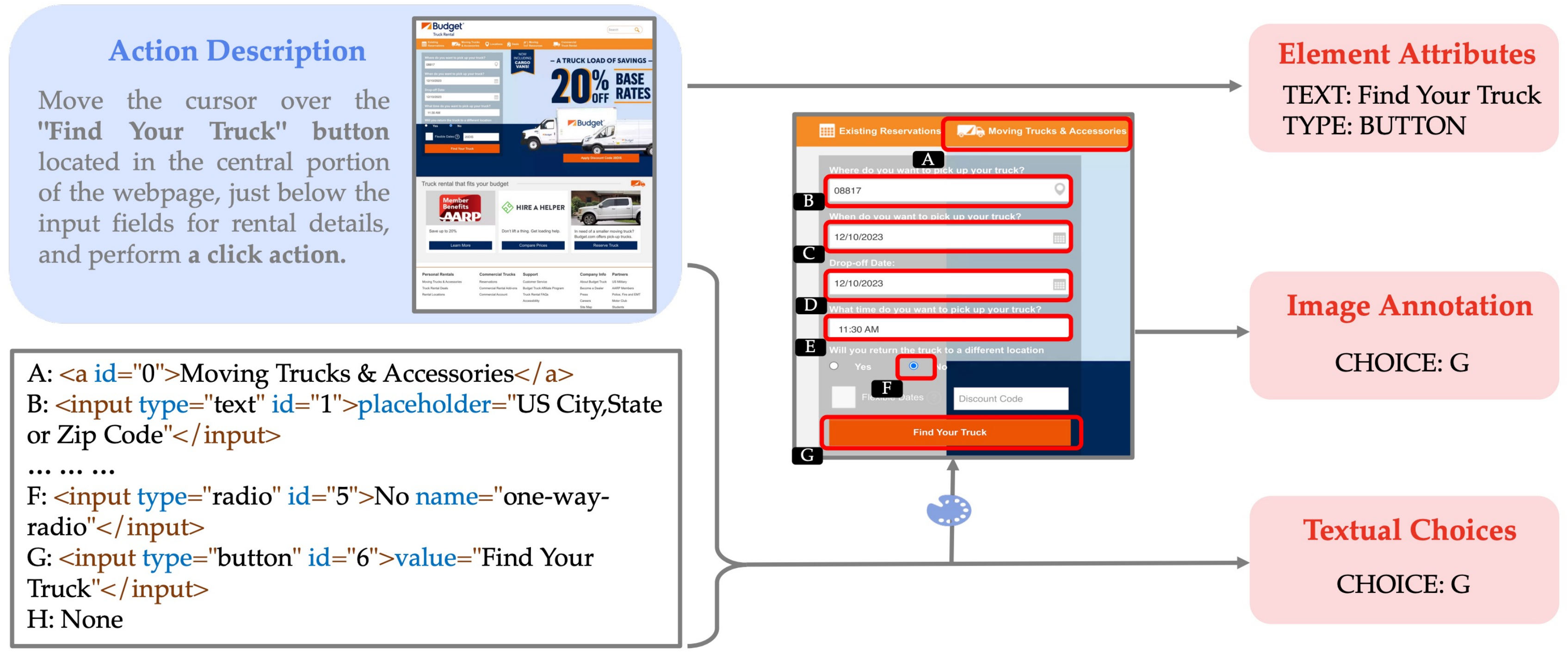
Embodiment of GUI Agents Before SeeAct-V



Mind2Web [NeurIPS'23 Spotlight]


Rent a truck near zip 08817 on December 10 at 11:30AM returned to the exact location and date.
LMM
Move the cursor over the "Find Your Truck" button located in the central portion of the webpage, just below the input fields for rental details, and perform a click action.
Grounding
Element: <input id=19 button find your truck />
Operation: CLICK
SeeAct
Browser Event

SeeAct [ICML'24]



Action Description

Move the cursor over the "Find Your Truck" button located in the central portion of the webpage, just below the input fields for rental details, and perform **a click action.**

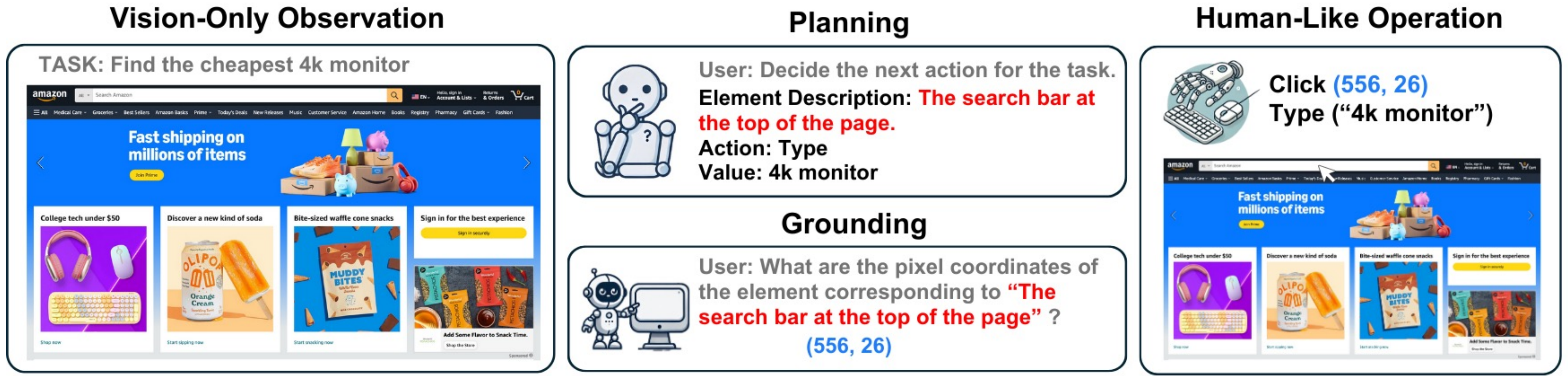
A: <a id="0">Moving Trucks & Accessories</a>
B: <input type="text" id="1">placeholder="US City,State or Zip Code"</input>
… … …
F: <input type="radio" id="5">No name="one-way-radio"</input>
G: <input type="button" id="6">value="Find Your Truck"</input>
H: None

Element Attributes
TEXT: Find Your Truck
TYPE: BUTTON

Image Annotation
CHOICE: G

Textual Choices
CHOICE: G

SeeAct-V: Human-Like, Vision-Centric Agents



Vision-Only Observation

TASK: Find the cheapest 4k monitor

Planning

User: Decide the next action for the task.
Element Description: The search bar at the top of the page.
Action: Type
Value: 4k monitor

Grounding

User: What are the pixel coordinates of the element corresponding to "The search bar at the top of the page" ?
(556, 26)

Human-Like Operation

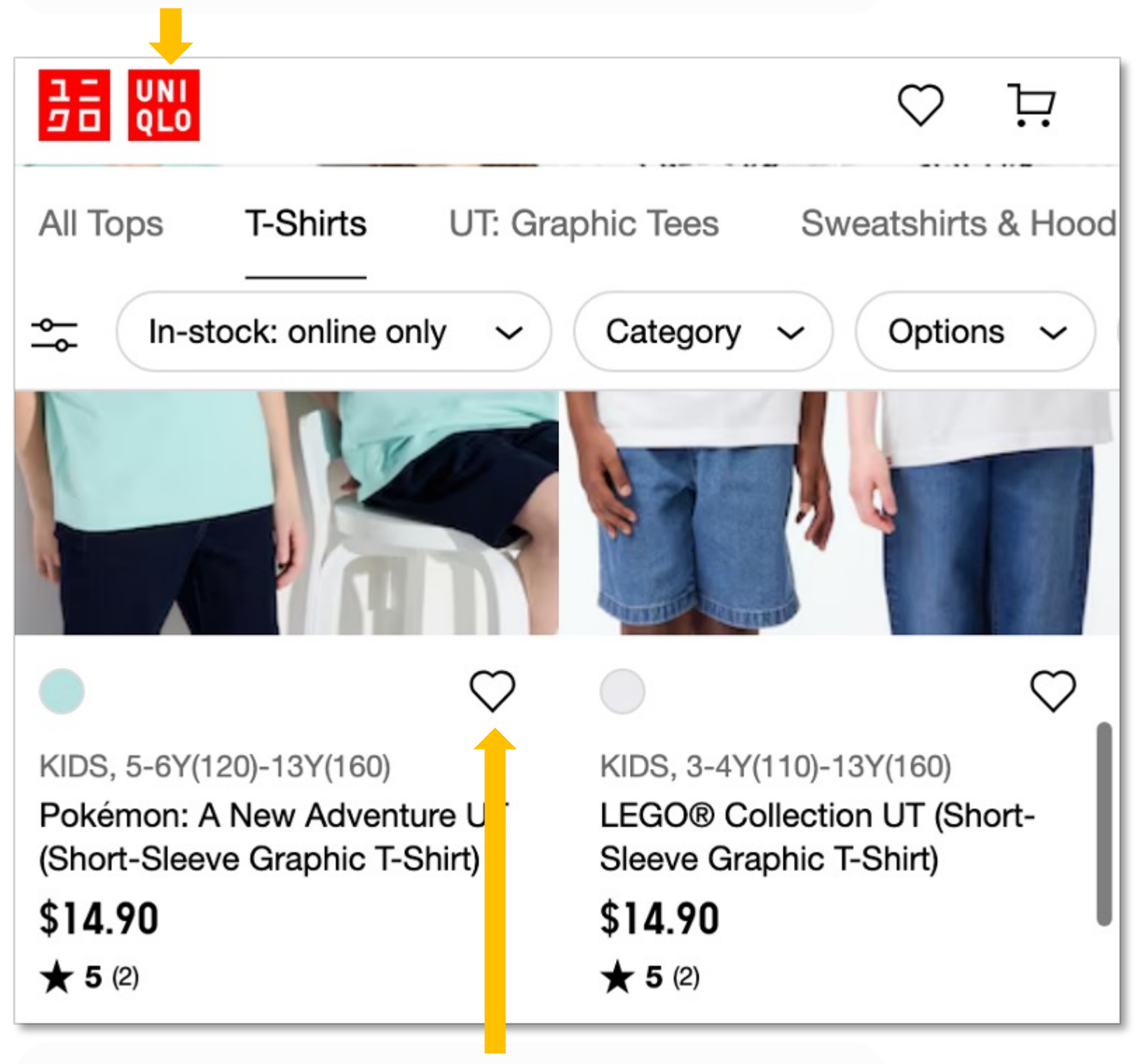
Click (556, 26)
Type ("4k monitor")

	Mind2Web (23' May)	SeeAct (24' Jan)	SeeAct-V (24' Aug)
Sensory Inputs	HTML/DOM	Screenshot + DOM	Screenshot Only
Effectors	Multi-choice Selection	Multi-choice Selection	Pixel-level Operations

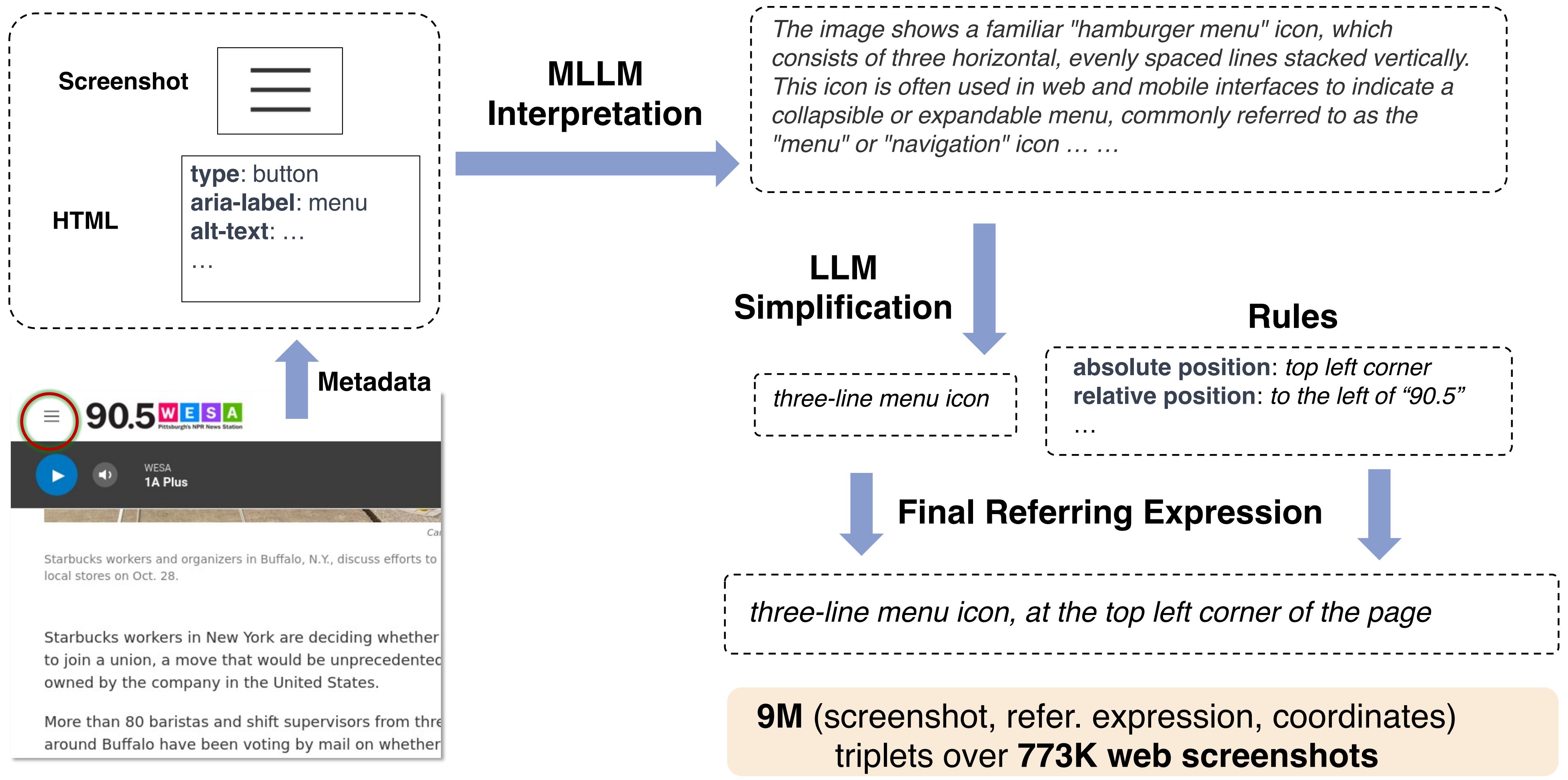
Web-Based Synthetic Data for Universal Visual Grounding

Universal: A model generalizes across various **referring expressions** (visual, positional, functional, hybrid) and all **platforms** (web, mobile, desktop)

1. Red icon labeled "UNIQLO"
2. Button at the top left corner
3. Navigate back to the homepage



1. Hollow heart button
2. Button below the Pokémon shirt
3. Favor the Pokémon shirt



type: button
aria-label: menu
alt-text: …
…

The image shows a familiar "hamburger menu" icon, which consists of three horizontal, evenly spaced lines stacked vertically. This icon is often used in web and mobile interfaces to indicate a collapsible or expandable menu, commonly referred to as the "menu" or "navigation" icon … …

three-line menu icon

absolute position: *top left corner*
relative position: *to the left of "90.5"*
…

three-line menu icon, at the top left corner of the page

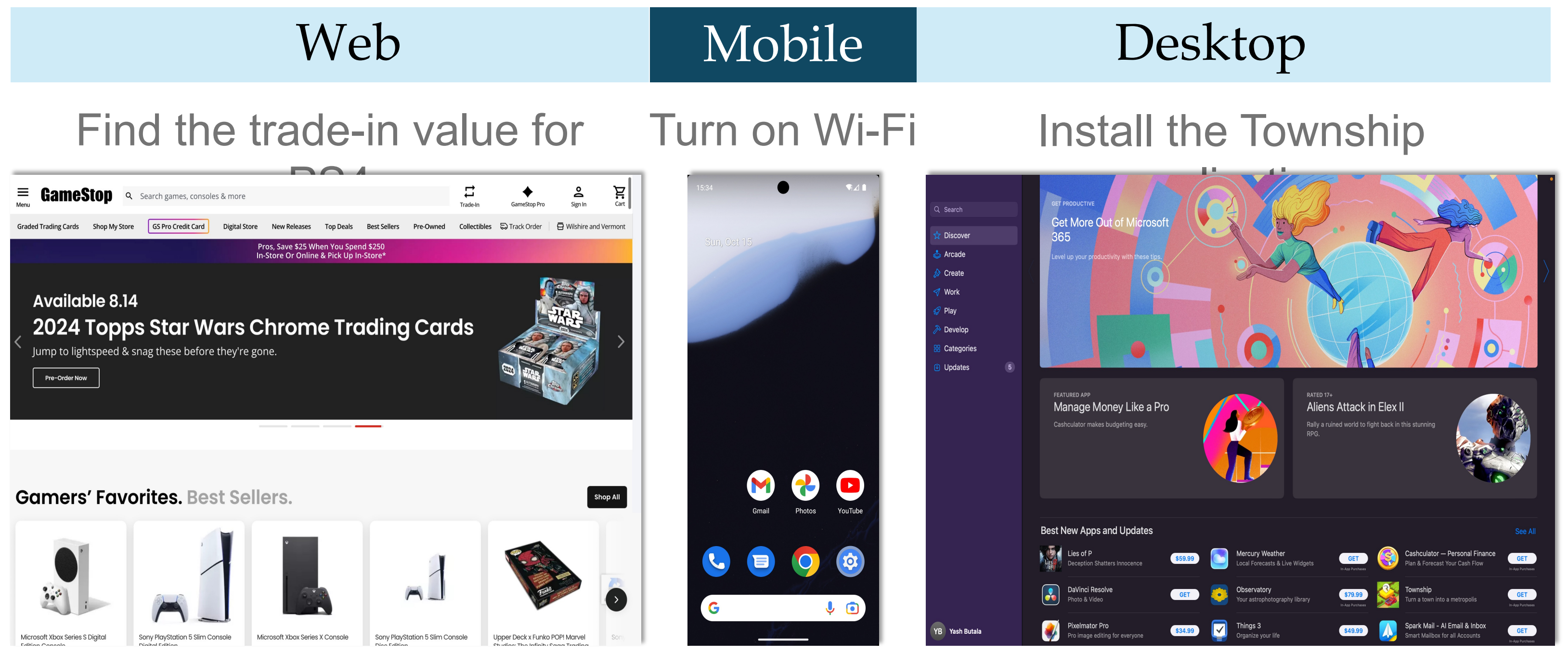
Starbucks workers and organizers in Buffalo, N.Y., discuss efforts to local stores on Oct. 28.

Starbucks workers in New York are deciding whether to join a union, a move that would be unprecedented owned by the company in the United States.

More than 80 baristas and shift supervisors from thre around Buffalo have been voting by mail on whether

9M (screenshot, refer. expression, coordinates) triplets over **773K web screenshots**

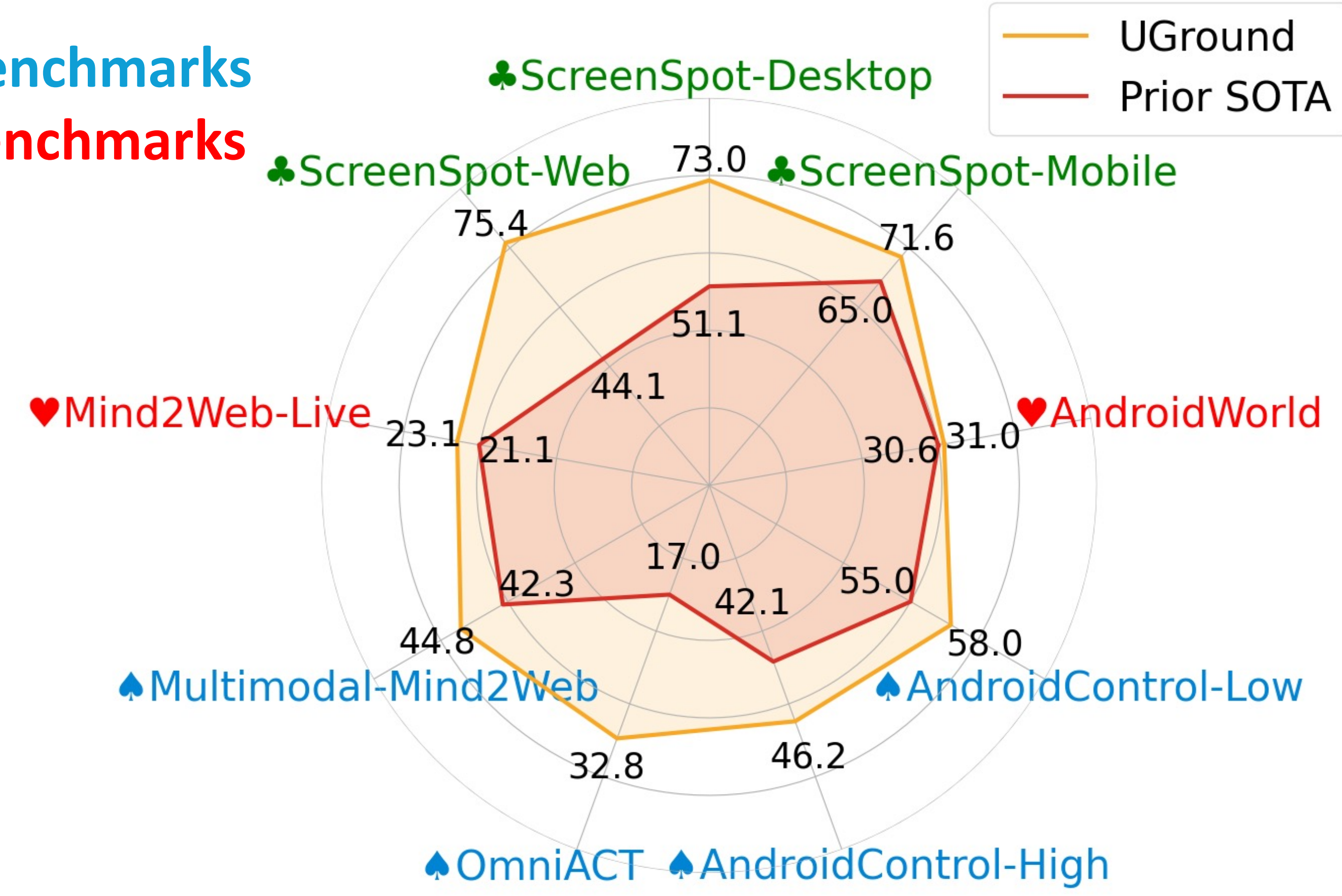
Minimal Design, SOTA Results


Web
Mobile
Desktop
Find the trade-in value for PS4
Turn on Wi-Fi
Install the Township
GameStop
Available 8.14
2024 Topps Star Wars Chrome Trading Cards
Gamers' Favorites. Best Sellers.

Most Comprehensive Evaluation (Setting Today's Standard)

(1) GUI Grounding
(2) Offline Agent Benchmarks
(3) Online Agent Benchmarks

UGround
+
SeeAct-V
=
SOTA Across the Board



Updates: Same Data x Qwen2-VL

Same Data (95% Web, 5% Android, 0% Desktop) + **Qwen2–VL** (2B, 7B, 72B)

	ScreenSpot	Mobile	Desktop	Web	Avg
	GPT-4o (OpenAI)	22.6	22.4	10.0	18.3
	Ferret-UI-Llama-8b (Apple)	48.4	28.7	20.0	32.3
	CogAgent (Zhipu)	45.5	47.1	49.5	47.4
	SeeClick	65.0	51.1	44.1	53.4
	OmniParser (Microsoft)	75.5	77.5	66.2	73.0
▶	**UGround (Initial)**	71.6	73.1	75.4	73.3
	ShowUI	83.9	68.7	72.7	75.1
	Molmo-7B-D (AI2)	77.2	75.0	73.4	75.2
▶	**UGround-V1-2B**	80.7	77.2	75.1	77.7
	Molmo-72B (AI2)	86.1	75.2	74.5	78.6
	OS-Atlas-Base-7B (Shanghai AI Lab)	83.0	77.4	82.6	81.0
	Aria-UI	83.1	78.8	81.4	81.1
	Claude-Computer-Use (Anthropic)	**91.9**	68.5	88.3	82.9
	Aguvis-7B	86.7	80.5	81.8	83.0
	Project Mariner (Google)				84.0
	CogAgent-9B (Zhipu)				85.4
▶	**UGround-V1-7B**	86.5	85.1	87.5	86.3
	Aguvis-72B	89.9	86.7	88.6	88.4
▶	**UGround-V1-72B**	88.8	**90.3**	**89.2**	**89.4**

⬅ **ScreenSpot**

UGround sparked the current wave of **UI-grounding research** and remains the **most effective open-source synthetic dataset.**

UGround continues to lead on **ScreenSpot** and recent desktop-centric benchmarks (**ScreenSpot-Pro**, **UI-Vision**) despite **using no desktop data.**

Check more results and resources on our homepage!