Less is More: Empowering GUI Agent with Context-Aware Simplification - Appendix -

Gongwei Chen[1], Xurui Zhou[1], Rui Shao[1]†, Yibo Lyu[1], Kaiwen Zhou[2],
Shuai Wang[2], Wentao Li[2], Yinchuan Li[2], Zhongang Qi[2], Liqiang Nie[1]†
[1]Harbin Institute of Technology, Shenzhen [2]Huawei Noah's Ark Lab
chengongwei@hit.edu.cn shaorui@hit.edu.cn nieliqiang@gmail.com
https://github.com/JiuTian-VL/SimpAgent

A. GUI navigation tasks

Android In The Wild (AITW) [11] consists of 30k instructions and 715k operation trajectories in the context of smartphone environments. However, the original train-test split carries the risk of overfitting due to overlapping instructions between the training and test sets, as well as the presence of multiple similar trajectories per instruction [2]. Therefore, we adopt the instruction-wise split scheme proposed by SeeClick [2] for the AITW dataset to achieve a fair evaluation. we follow the same data processing settings in SeeClick. The action space of AITW consists of 12 actions: `CLICK`, `TYPE`, `SELECT`, `SCROLL UP`, `SCROLL DOWN`, `SCROLL LEFT`, `SCROLL RIGHT`, `PRESS BACK`, `PRESS HOME`, `PRESS ENTER`, `STATUS TASK COMPLETE`, `STATUS TASK IMPOSSIBLE`.

Mind2Web [3] comprises over 2000 open-ended tasks collected from 137 real websites, originally for text-based web agents, provides only HTML observations. To support visual agents, we extracted screenshots and target bounding boxes. Since full-page captures (e.g. 1920×12000) are impractical for LVLMs, we follow the same data processing settings in SeeClick, cropping around target elements, standardizing the resolution to 1920×1080. The action space of Mind2Web consists of 2 actions: `CLICK`, `TYPE`.

GUI-Odyssey [9] is a comprehensive dataset for training and evaluating cross-app navigation agents, consisting of 7,735 episodes from 6 mobile devices. We follow the same data processing and evaluation settings in GUI-Odyssey. The action space of GUI-Odyssey consists of 9 actions: `CLICK`, `SCROLL`, `LONG PRESS`, `TYPE`, `COMPLETE`, `IMPOSSIBLE`, `HOME`, `BACK`, `RECENT`.

AndroidControl [7] encompasses 14,548 unique tasks across 833 Android apps, offering both high and low-level instructions to explore the level of task complexity an agent can handle. We follow the same data processing and evaluation settings in OS-Atlas [13]. The action space of AndroidControl consists of 9 actions: `CLICK`, `SCROLL`, `LONG PRESS`, `TYPE`, `NAVIGATE HOME`, `NAVIGATE BACK`, `OPEN APP`, `WAIT`, `TERMINATE`.

B. Implementation Details

We use Qwen2VL-2B as our base model. For different datasets, we adopt the same data organization format following SeeClick [2]. We parse the executed actions in JSON format, dividing them into two parts: action_type and action_value. Next, we introduce the format and examples of the training and testing data.

For the 4A data format, we use the same prompt in SeeClick to execute each step of the agent.

```
"<image>Please generate the next move according to
the ui screenshot, instruction and previous actions.
Instruction: What's on the menu at Domino's?.
Previous actions:
Step0: {\"action_type\": PRESS HOME}.
Step1: {\"action_type\": CLICK, \"click_point\":
(524,865)}.
Step2: {\"action_type\": TYPE, \"typed_text\":
\"menu at Domino's\"}.
Step3: {\"action_type\": CLICK, \"click_point\":
(325,156)}. "
```

For the 4AO data format, we use the following prompt to execute each step of the agent.

```
"Please generate the next move according to the
instruction, previous actions, previous ui screenshot
and current ui screenshot.
Instruction: What's on the menu at Domino's?.
Image_0:<image>
Step_0:{\"action_type\": PRESS HOME} .
Image_1:<image>
Step_1:{\"action_type\": CLICK, \"click_point\":
(524,865)} .
Image_2:<image>
Step_2:{\"action_type\": TYPE, \"typed_text\": \"menu
at Domino's\"} .
Image_3:<image>
Step_3:{\"action_type\": CLICK, \"click_point\":
(325,156)} .
Image_4:<image>"
```

†Corresponding authors

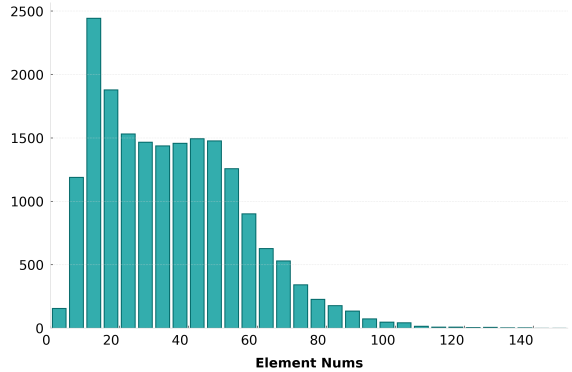


Figure 1. The distribution of element nums in AITW dataset.

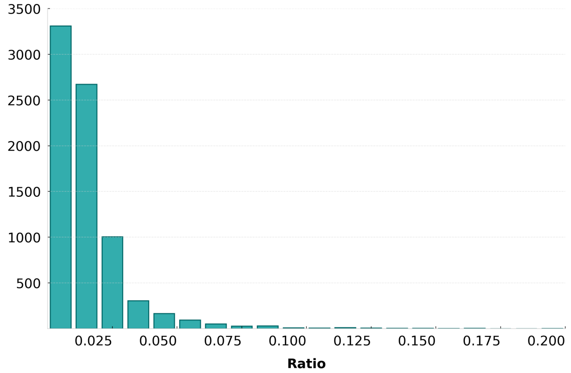


Figure 2. The distribution of target element ratio in Mind2Web dataset.

For the AITW [11] and GUI-Odyssey [9] datasets, we employ four low-resolution historical images, adhering to the 4AO format proposed by SeeClick. Each image's longest side is scaled to 512 pixels, maintaining the aspect ratio and ensuring the resolution stayed within 512x512.

For AndroidControl [7] and Mind2Web [3] datasets, we adopt the same setting in ShowUI, retaining two high-resolution historical images, with a maximum of 1280 tokens. We choose not to use four due to GPU memory constraints.

We use bfloat16 precision for training. All experiments are conducted with LoRA fine-tuning, applying a rank of 8 and an alpha value of 16 exclusively for the language model. We leverage DeepSpeed Zero-2 and utilize flash attention to accelerate training. For AITW and GUI-Odyssey, we use the same training strategy in SeeClick, with a learning rate of 3e-5 and a global batch size of 64. For AndroidControl, we use a learning rate of 3e-4 and a global batch size of 128. For Mind2Web, we use a learning rate of 3e-4 and a global batch size of 16.

C. The statistical Analyses of GUI Navigation Datasets

In Figure 1, We present the distribution of the element nums included in screenshots from the AITW dataset. The number of elements per screenshot typically falls within the range of 10 to 100, reflecting the characteristic of GUI scenes having a large number of visual elements.

In Figure 2, we present the distribution of the target element ratio in the Mind2Web dataset. It is evident that most elements occupy less than 3% of the entire screen, a characteristic that contributes to the difficulty in element localization in GUI scenes.

D. More detailed results

In Tables 1, 2, 3, we present the performances on all subsets of each navigation dataset. On AITW (6.5 steps on average), our SimpAgent significantly outperforms ShowUI, except the “Single” subset (1.5 steps on average). We attribute it to the large-scale grounding pre-training of ShowUI, which enhances its single-horizon task completion ability. On GUI-Odyssey (15.4 steps on average), SimpAgent achieves the performance gains of 3.41% on the “Multi-Apps” subset (22.3 steps on average), 3.58% on the “Shopping” subset (18.2 steps on average), demonstrating its superior ability of accomplishing long-horizon tasks. On Mind2Web, our SimpAgent and SimpAgent-M produce new state-of-the-art performances on three out-of-domain test sets. Our SimpAgent-M is slightly inferior to SimpAgent and comparable with the baseline Qwen2VL on the “Cross-Website” test set, and the performance gap is relatively small. We speculate that the insignificant performance difference may be due to the limited size of the “Cross-Website” set (only 177 tasks), which is insufficient to yield reliable experimental conclusions.

E. Case Study

In Figure 3, we present more navigation tasks. The baseline model (fine-tuned Qwen2-VL) is unable to complete these steps. In contrast, the visualizations show that our SimpAgent can effectively reduce interference and correctly identify relevant elements while ignoring irrelevant ones. In Figure 4, we present more cases about the attention difference between w/ and w/o consistency guidance. These visualizations demonstrate that the historical action tokens pay more attention to the adjacent historical vision tokens under the setting of consistency guidance, facilitating a better information aggregation. In Figure 5 and 6, we present the complete task execution trajectory of SimpAgent.

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A Frontier Large Vision-Language Model with Versatile Abilities. *arXiv preprint*, 2023. 3, 4

Method	#P.T.	#Param.	General	Install	GoogleApps	Single	WebShopping	Overall
GPT-4V [10]	-	-	41.7	42.6	49.8	72.8	45.7	50.5
QwenVL [1]	-	9.6B	49.5	59.9	46.9	64.7	50.7	54.3
SeeClick [2]	850K	9.6B	54.0	66.4	54.9	63.5	57.6	59.3
Qwen2VL* [12]	-	2B	59.4	71.7	64.2	71.8	63.0	66.0
Qwen2VL [12]	-	2B	61.2	73.5	72.1	73.8	64.2	69.0
Iris [4]	850K	9.6B	61.5	71.4	58.3	66.4	60.2	63.6
ShowUI [8]	256K	2B	63.9	72.5	69.7	**77.5**	66.6	70.0
SimpAgent	-	2B	**64.3**	75.5	73.7	75.8	67.0	71.3
SimpAgent-M	-	2B	64.1	**75.8**	**74.0**	76.2	**67.2**	**71.5**

Table 1. Step success rate on AITW, divided by domains. “#P.T.” and “#Param.” denote the number of pre-training GUI datasets and parameters, respectively. “*” means the variant utilizing only action history. “-M” denotes only applying masking-based element pruning without inference FLOPs reduction.

Method	#Param.	Cross-Task			Cross-Website			Cross-Domain		
		Ele.Acc	Op.F1	Step SR	Ele.Acc	Op.F1	Step SR	Ele.Acc	Op.F1	Step SR
Qwen-VL [1]	9.6B	15.9	86.7	13.3	13.2	83.5	9.2	14.1	84.3	12.0
CogAgent [5]	18B	22.4	53.0	17.6	18.4	42.4	13.4	20.6	42.0	15.5
SeeClick [2]	9.6B	28.3	87.0	25.5	21.4	80.6	16.4	23.2	84.8	20.8
Qwen2VL* [12]	2B	49.7	89.1	45.0	46.8	**86.5**	40.9	45.4	86.6	40.5
Qwen2VL [12]	2B	51.6	88.6	46.7	48.5	85.7	42.2	48.3	87.0	44.6
Iris [4]	9.6B	33.5	87.1	32.0	31.2	82.2	26.2	32.8	85.1	28.8
ShowUI [8]	2B	39.9	88.6	37.2	41.6	83.5	35.1	39.4	86.8	35.2
SimpAgent	2B	51.0	89.2	47.1	**48.7**	86.0	**42.8**	46.9	86.5	43.3
SimpAgent-M	2B	**52.4**	**89.4**	**48.7**	48.2	85.8	42.2	**49.0**	**88.2**	**45.0**

Table 2. Performance comparison on Mind2Web across different settings. We report element accuracy (Ele.Acc), operation F1 (Op.F1), and step success rate (Step SR). “#Param.” denotes the number of parameters. “*” means the variant utilizing only action history. “-M” denotes only applying masking-based element pruning without inference FLOPs reduction.

[2] Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Yantao Li, Jianbing Zhang, and Zhiyong Wu. SeeClick: Harnessing GUI Grounding for Advanced Visual GUI Agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. arXiv, 2024. 1, 3

[3] Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2Web: Towards a Generalist Agent for the Web. In *Advances in Neural Information Processing Systems, 2023*. arXiv, 2023. 1, 2

[4] Zhiqi Ge, Juncheng Li, Xinglei Pang, Minghe Gao, Kaihang Pan, Wang Lin, Hao Fei, Wenqiao Zhang, Siliang Tang, and Yueting Zhuang. Iris: Breaking GUI Complexity with Adaptive Focus and Self-Refining. *arXiv preprint*, 2024. 3

[5] Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxuan Zhang, Juanzi Li, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. CogAgent: A Visual Language Model for GUI Agents. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. arXiv, 2023. 3, 4

[6] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint*, 2024. 4

[7] Wei Li, William Bishop, Alice Li, Chris Rawles, Folawiyo Campbell-Ajala, Divya Tyamagundlu, and Oriana Riva. On the Effects of Data Scale on Computer Control Agents. In *Advances in Neural Information Processing Systems*. arXiv, 2024. 1, 2

[8] Kevin Qinghong Lin, Linjie Li, Difei Gao, Zhengyuan Yang, Shiwei Wu, Zechen Bai, Weixian Lei, Lijuan Wang, and Mike Zheng Shou. ShowUI: One Vision-Language-Action Model for GUI Visual Agent. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3

[9] Quanfeng Lu, Wenqi Shao, Zitao Liu, Fanqing Meng, Boxuan Li, Botong Chen, Siyuan Huang, Kaipeng Zhang, Yu Qiao, and Ping Luo. GUI Odyssey: A Comprehensive Dataset for Cross-App GUI Navigation on Mobile Devices. *arXiv preprint*, 2024. 1, 2, 4

[10] OpenAI. Gpt-4 technical report, 2023. 3, 4

[11] Christopher Rawles, Alice Li, Daniel Rodriguez, Oriana Riva, and Timothy Lillicrap. Android in the Wild: A Large-Scale Dataset for Android Device Control. In *Advances in Neural Information Processing Systems, 2023*, 2023. 1, 2

[12] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Jun-

Method	#P.T.	#Param.	Tool	Information	Shopping	Media	Social	Multi-Apps	Overall
GPT-4V [10]	-	-	23.49	20.16	19.15	16.92	13.83	19.02	18.76
GPT-4o [6]	-	-	20.81	16.28	31.91	15.38	21.28	16.67	20.39
CogAgent [5]	140M	18B	15.66	10.74	9.15	11.66	13.08	10.73	11.84
QwenVL [1]	-	9.6B	83.11	65.70	62.43	76.38	76.12	73.10	72.81
Qwen2VL* [12]	-	2B	80.35	62.42	59.52	71.11	72.75	67.91	69.01
Qwen2VL [12]	-	2B	83.71	67.44	64.79	77.19	79.08	76.97	74.86
OdysseyAgent [9]	-	9.6B	85.16	68.53	62.87	76.49	77.61	74.83	74.25
SimpAgent	-	2B	84.91	69.89	66.45	78.05	79.21	78.24	76.02
SimpAgent-M	-	2B	**86.57**	**70.08**	**67.65**	**79.78**	**81.04**	**79.21**	**77.39**

Table 3. Step success rate on GUI-Odyssey Test-Random split. “#P.T.” and “#Param.” denote the number of pre-training GUI datasets and parameters, respectively. “*” means the variant utilizing only action history. “-M” denotes only applying masking-based element pruning without inference FLOPs reduction.

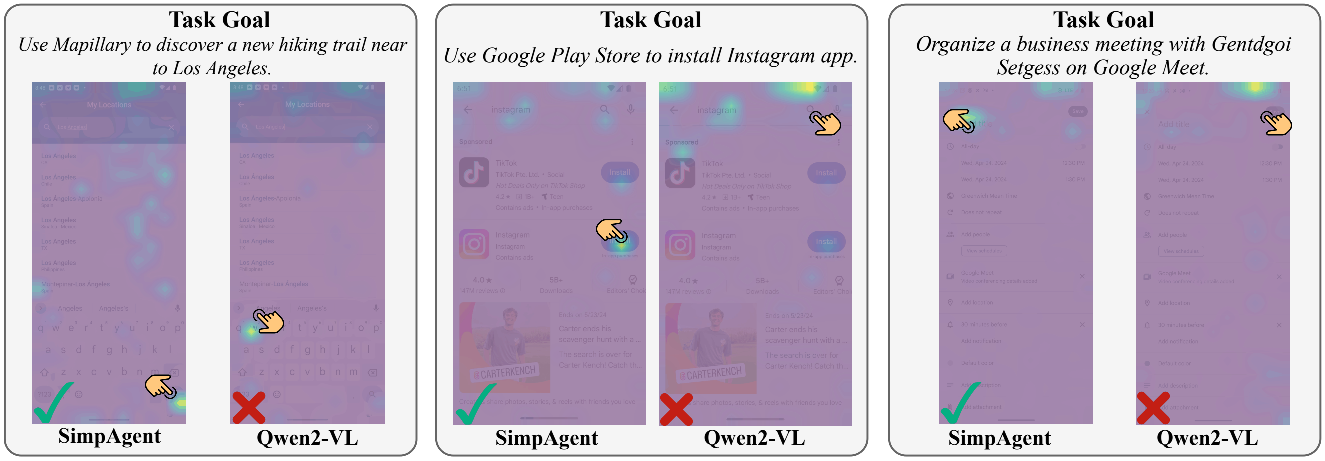

Task Goal
Use Mapillary to discover a new hiking trail near to Los Angeles.
SimpAgent
Qwen2-VL
Task Goal
Use Google Play Store to install Instagram app.
SimpAgent
Qwen2-VL
Task Goal
Organize a business meeting with Gentdgoi Setgess on Google Meet.
SimpAgent
Qwen2-VL

Figure 3. Illustration of navigation steps in the GUI-Odyssey dataset. SimpAgent distinguishes the correct element among various confusing elements. This demonstrates the effectiveness of our proposed Masking-based Element Pruning method.

yang Lin. Qwen2-VL: Enhancing Vision-Language Model’s Perception of the World at Any Resolution. *arXiv preprint*, 2024. 3, 4

[13] Zhiyong Wu, Zhenyu Wu, Fangzhi Xu, Yian Wang, Qiushi Sun, Chengyou Jia, Kanzhi Cheng, Zichen Ding, Liheng Chen, Paul Pu Liang, and Yu Qiao. OS-ATLAS: A Foundation Action Model for Generalist GUI Agents. *arXiv preprint*, 2024. 1

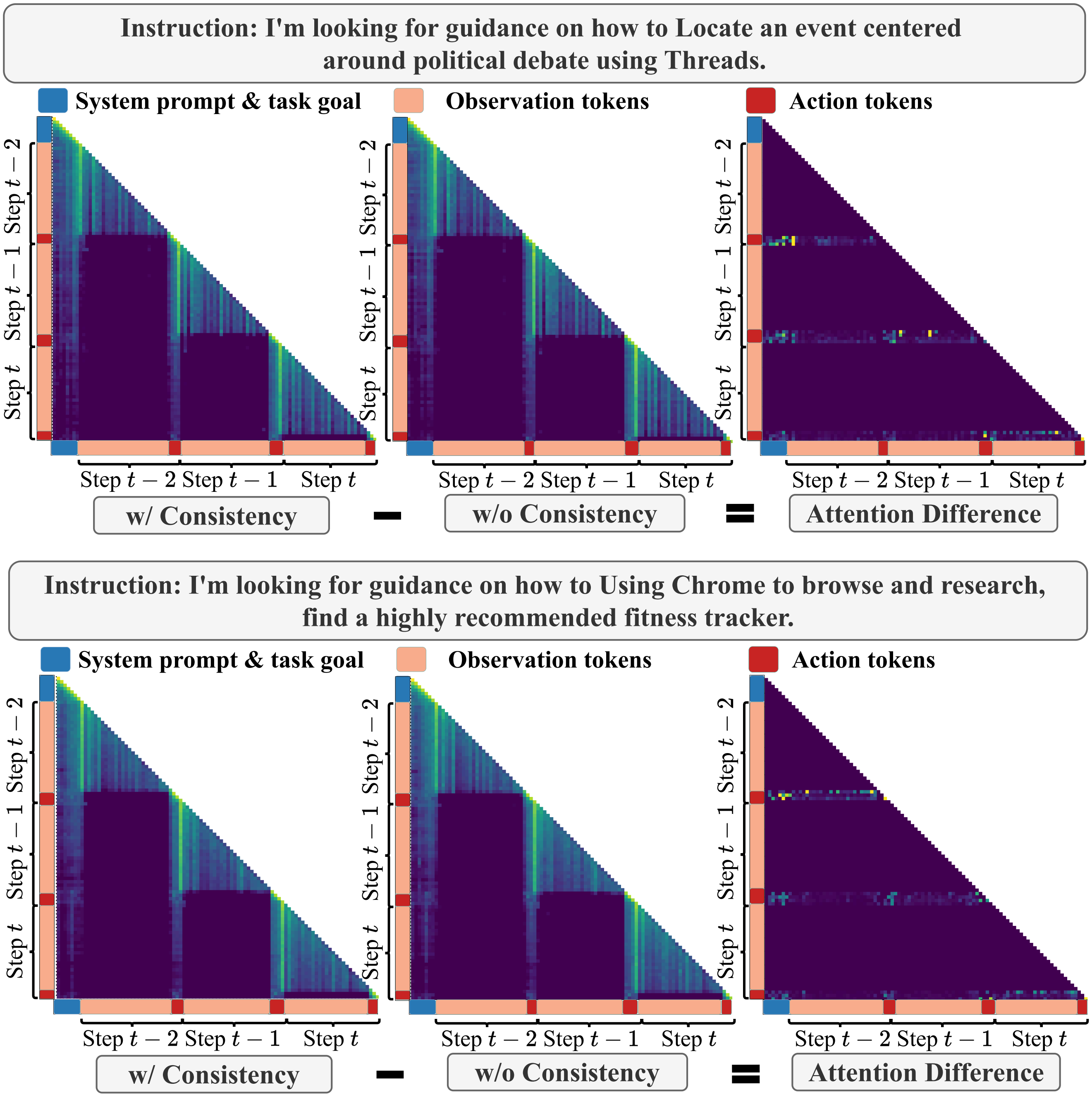
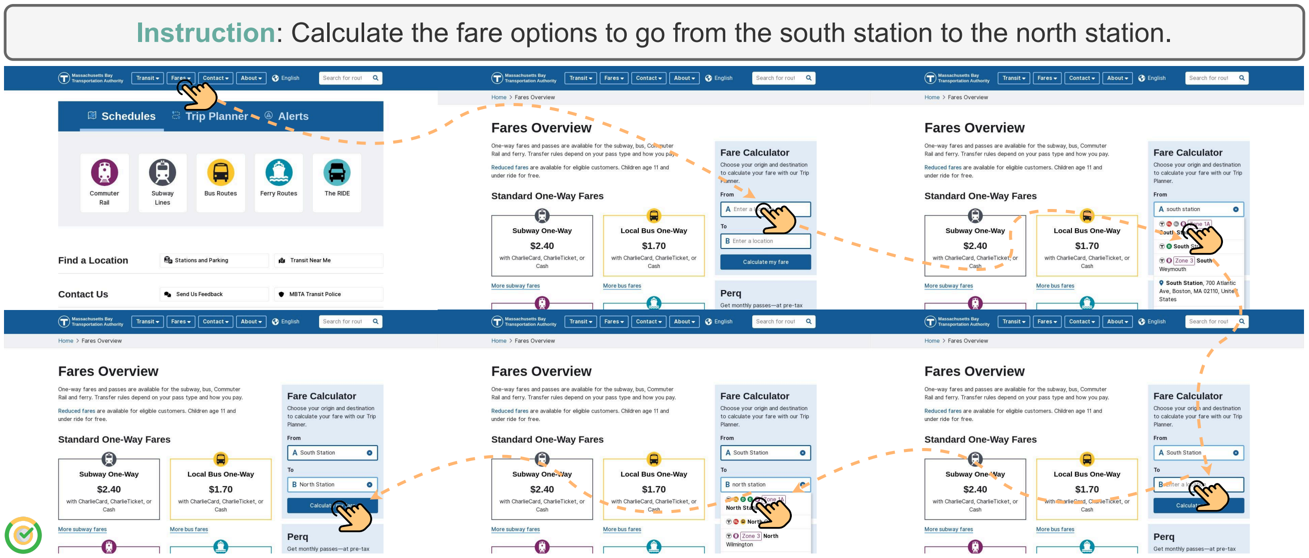
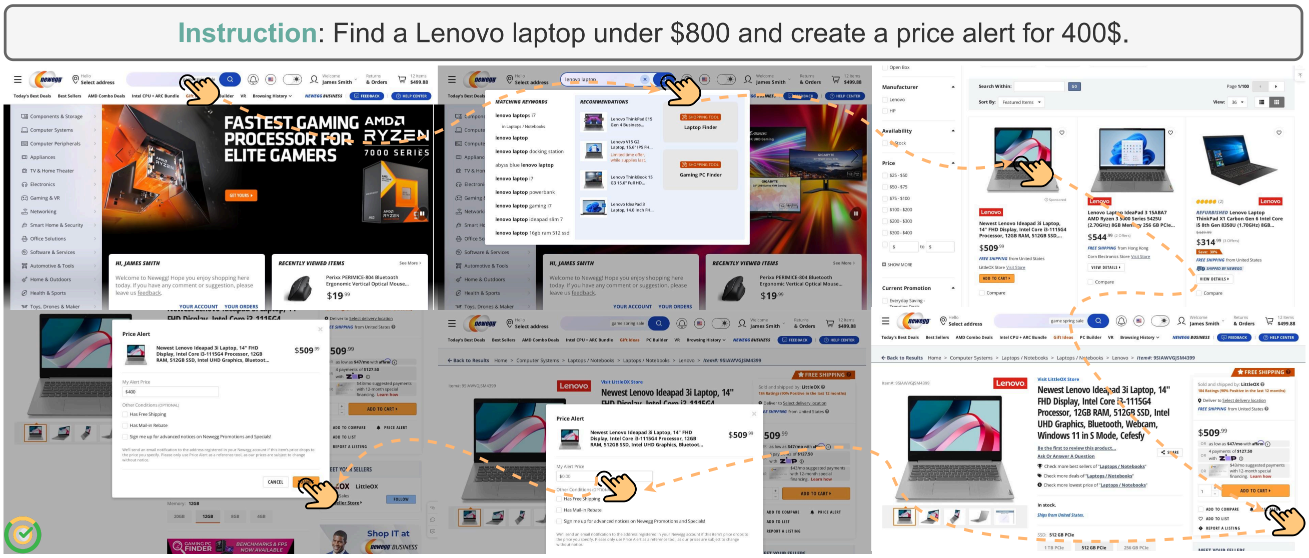


Figure 4. Illustration of attention maps in agent models w/ and w/o consistency guidance, and their attention difference map. The **attention difference map** shows that action tokens pay more attention (highlighted positions) to historical observation tokens when they act as query tokens with consistency guidance. This attention comparison demonstrates that consistency guidance can promote the information aggregation from observations to actions and facilitate the history compression.



Instruction: Calculate the fare options to go from the south station to the north station.

Figure 5. The real-world application of SimpAgent when adapted to downstream GUI navigation tasks.



Instruction: Find a Lenovo laptop under $800 and create a price alert for 400$.

lenovo laptop
FASTEST GAMING PROCESSOR FOR ELITE GAMERS
AMD RYZEN 7000 SERIES
Lenovo
$509.99
$544.99
$314.99
Price Alert
Newest Lenovo Ideapad 3i Laptop, 14" FHD Display, Intel Core i3-1115G4 Processor, 12GB RAM, 512GB SSD, Intel UHD Graphics, Bluetoot...
My Alert Price
$400
CANCEL
ADD TO CART

Figure 6. The real-world application of SimpAgent when adapted to downstream GUI navigation tasks.