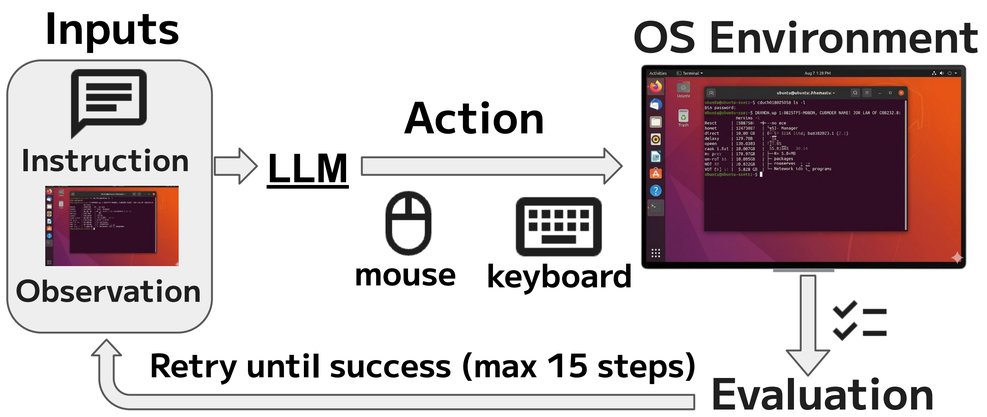
A11y-Compressor: A Framework for Enhancing the Efficiency of GUI Agent Observations through Visual Context Reconstruction and Redundancy Reduction

Michito Takeshita, Takuro Kawada, Takumi Ohashi, Shunsuke Kitada, Hitoshi Iyatomi
Hosei University, Tokyo, Japan

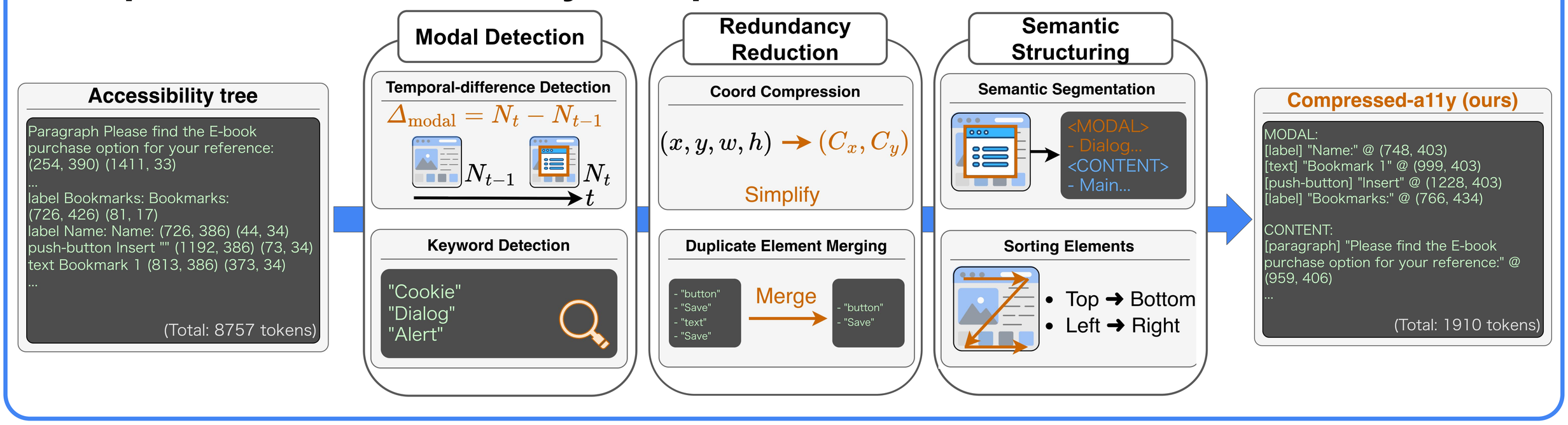
ACL 2026 SAN DIEGO JULY 2-7

Motivation

- AI agents that operate GUIs require effective observation representations for reliable grounding [Gou+, ICLR'25; Xie+, NeurIPS'24]
- Accessibility (A11y) trees provide useful text-based observations, but they are redundant and poorly aligned with GUI visual structure
- To address these limitations, we introduce A11y-Compressor, a framework for compact, structured GUI observations from A11y trees



Proposed framework: A11y-Compressor



Evaluation

- **Evaluation Benchmark:** OSWorld [Xie+, NeurIPS'24]

  OSWorld evaluates GUI agents on real-desktop operations across 10 application domains, such as web browsing in Chrome, coding in VS Code, OS operations, and multi-app workflows.
- **Protocol:** Max 15 interaction steps per task; 2 trials per task; success if either trial succeeds
- **Observation representations:** A11y tree, Compressed-a11y (ours), LineRetriever* [Kerboua+ arXiv'25] based representation, Screenshot

  *LineRetriever : retrieves task-relevant A11y-tree lines with an LLM, extended beyond web tasks.
- **Model:** Qwen3-VL-32B [Bai+ arXiv'25] with recommended hyperparameters: temperature = 1.0, top_p = 0.9

Results

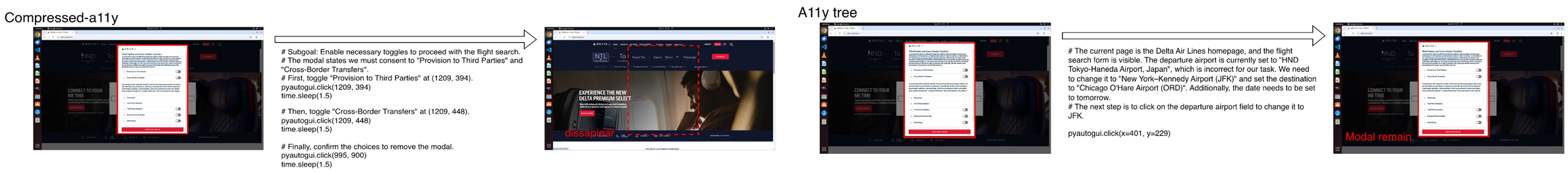
Token Efficiency: Compressed-a11y reduces observation tokens by 78% compared with A11y tree while preserving task-relevant UI information.

Method	Chrome	GIMP	Calc	Impress	Writer	Multi Apps	OS	Thunderbird	VLC	VS Code	Overall
A11y tree	5,538	3,352	30,181	5,585	4,110	8,607	820	3,040	1,609	7,228	8,758
LineRetriever	2,060	1,142	9,115	**1,634**	2,339	3,488	**381**	**1,709**	869	2,790	3,096
Compressed-a11y (ours)	**1,741**	**1,110**	**3,438**	1,859	**2,321**	**1,709**	536	2,045	**791**	**2,767**	**1,910**

Success Rate: Compressed-a11y achieves the best overall success rate and matches or outperforms baselines in most domains.

Method	Chrome	GIMP	Calc	Impress	Writer	Multi Apps	OS	Thunderbird	VLC	VS Code	Overall
Screenshot	0.045	0.115	0.000	0.021	0.043	**0.108**	0.208	0.000	0.118	0.043	0.070
A11y tree	0.182	0.192	0.000	0.149	0.087	**0.108**	0.333	0.267	**0.294**	0.304	0.156
LineRetriever	0.136	0.192	0.022	**0.191**	0.087	**0.108**	0.333	0.133	0.176	**0.348**	0.151
Compressed-a11y (ours)	**0.250**	**0.231**	**0.043**	**0.191**	**0.304**	**0.108**	**0.375**	**0.467**	**0.294**	**0.348**	**0.207**

Case Study: Compressed-a11y clarifies the interaction scope by separating active modal foregrounds from inactive background elements.



Future Work

- Incorporate visual cues such as icons, colors, and spatial layouts that are not captured in A11y tree.
- Extend evaluation beyond desktop applications to mobile interfaces and other GUI environments.