Dual-View Visual Contextualization for Web Navigation

Jihyung Kil, Chan Hee Song, Boyuan Zheng, Xiang Deng, Yu Su, Wei-Lun Chao
The Ohio State University

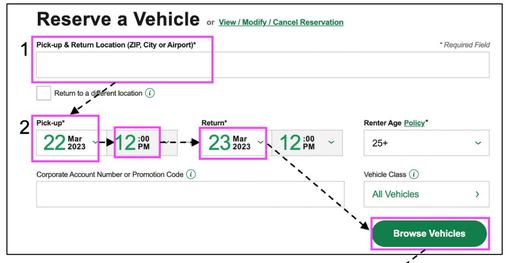
Highlights

- Propose **DUAL-VCR**, a simple and effective dual-view representation of HTML elements for web navigation.
- Web developers tend to arrange task-related elements nearby on webpages to enhance user experiences.
- **DUAL-VCR** thus contextualizes each HTML element with its neighbors in the webpage screenshot.
- **DUAL-VCR** consistently outperforms baselines on the real-world web navigation benchmark Mind2Web.

Web Navigation

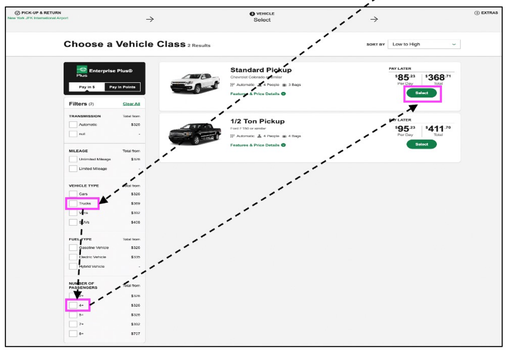
Task

Find the lowest rent truck for 4 people, pick up from JFK airport at 11 am on March 27 and return at noon on March 30.



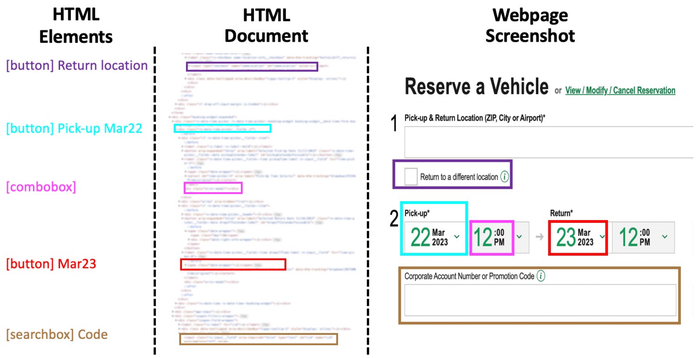
Action

HTML Element	Operation
[searchbox] Pick-up & Return	Type: JFK
[button] Pick-up	Click
[button] 03/27/2023	Click



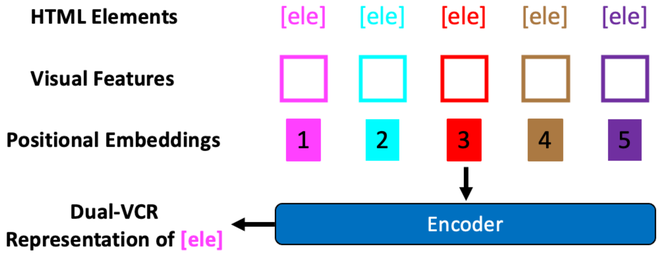
Overview

HTML elements can be found in both HTML document and webpage screenshot.



Neighboring HTML elements in the screenshot are semantically related and task-related to each other.

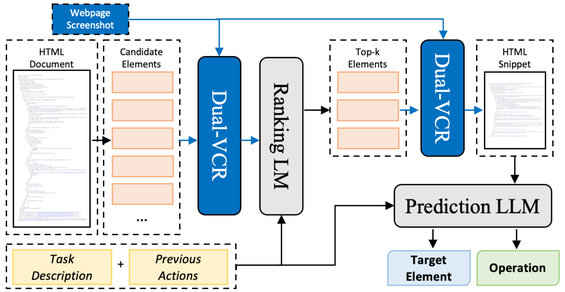
DUAL-VCR contextualizes each HTML element with its neighbors in the screenshot, using both textual and visual features.



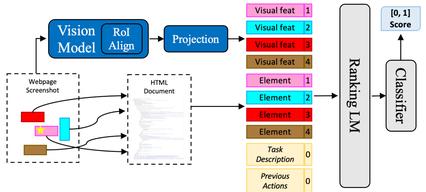
Dual-VCR Pipeline

There exists numerous HTML elements in the webpage. Directly feeding all elements into LLMs are infeasible or cost-prohibitive.

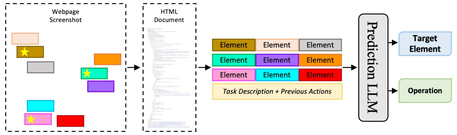
Use a ranker to identify important HTML elements for next action. Only pass Top-K elements into LLM for action prediction.



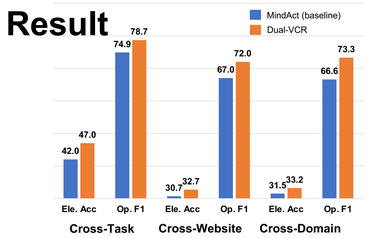
Dual-VCR Ranker



Dual-VCR Action Predictor



Result



Qualitative Result

t=4

GT: [combobox] Select 11:00 am
MindAct: [button] Vehicles Click
Dual-VCR: [combobox] Select 11:00 am