OmniParser for Pure Vision-Based GUI Agent

Microsoft Research

January 2025

Introduction

The recent success of large vision language models shows great potential in driving the agent system operating on user interfaces. The power of multimodal models like GPT-4V/o as a general agent on multiple operating systems across different applications is largely underestimated, due to the lack of a robust screen parsing technique capable of: 1. reliably identifying interactable icons within the user interface, and 2. understanding the semantics of various elements in a screenshot and accurately associating the intended action with the corresponding region on the screen.

Meet OmniParser, a compact screen parsing module that converts UI screenshots into structured elements. OmniParser can easily be paired with a variety of models to create agents capable of taking actions on UIs. When used with GPT-4V, it significantly improves the agent capability to generate precisely grounded actions for interface regions.

New Release Highlights

Version 1.5

- A newly finetuned interactive screen region detection model, that better captures more fine-grained icons, especially for higher resolution GUI screen and small icons.
- Capability of predicting interactivity of every parsed screen element.
- New icon dataset for Microsoft 365 applications.
- Improved logic for deduplication of the detected regions.

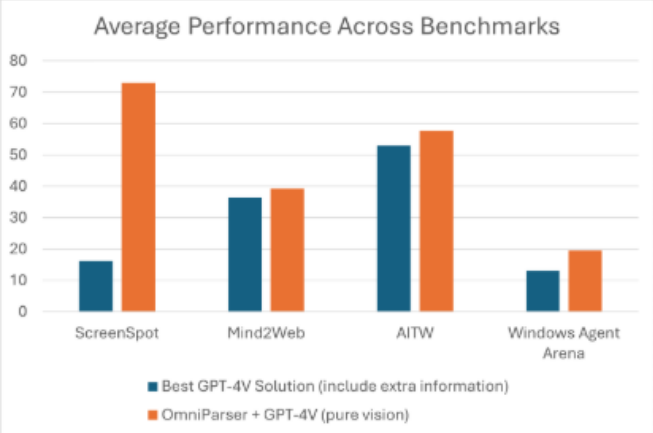
Key Features

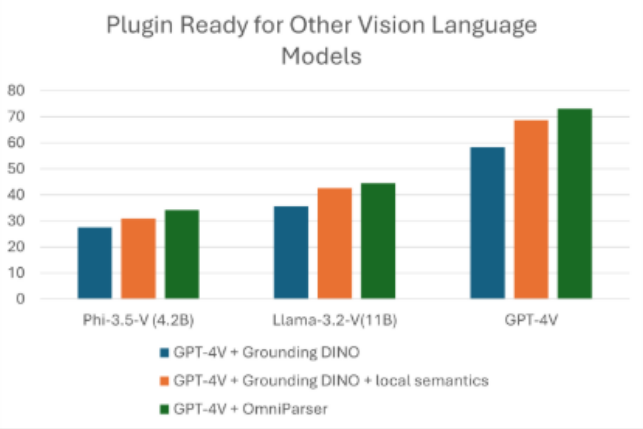
- **Interactive region detection:** we curated an interactable icon detection dataset from popular web pages with annotation to highlight clickable and actionable regions. An interactive region detection model is fine-tuned on the curated dataset, which reliably identifies actionable regions within a screenshot.

- **Icon functionality semantics.** We curated an icon description dataset, designed to associate each UI element with its corresponding function. We trained a light weight Florence2 model on the icon description dataset, which is used to extract the functional semantics of the detected elements, generating contextually accurate descriptions of their intended actions.

Benchmark results

We demonstrate that with the parsed results, the performance of GPT-4V is greatly improved on ScreenSpot benchmarks. On Mind2Web, OmniParser +GPT-4V achieves better performance compared to GPT-4V agent that uses extra information extracted from HTML. And on AITW benchmark, OmniParser outperforms GPT-4V augmented with specialized Android icon detection model that is trained with view hierarchy. It also achieves the best performance on the recently released Windows benchmark WindowsAgentArena!





Scenarios Where This Tech Shines

- **For users who need computer use agents but don't have resources to finetune their own model, OmniParser provides a plugin choice for off-the-shelf vision language models:** we show significant improved ScreenSpot benchmark performance of OmniParser combined with recently announced vision language models: Phi-3.5-V and Llama-3.2-V.
- **Building Cross platform + cross application GUI agents:** OmniParser serve as a general and easy-to-use tool that has the capability to parse general user screen across both PC and mobile platforms without any dependency on any extra information such as HTML and view hierarchy in Android.

Learn More

- Project Page; https://microsoft.github.io/OmniParser/ (5.4k stars)
- Model: https://huggingface.co/microsoft/OmniParser
- Paper: https://arxiv.org/pdf/2408.00203
- Blog: microsoft.com/en-us/research/articles/omniparser-for-pure-vision-based-gui-agent/
- Code: microsoft/OmniParser: A simple screen parsing tool towards pure vision based GUI agent
- Demo: OmniParser demo - a Hugging Face Space by microsoft

Contact Us

Project Contact: omniparser@microsoft.com