Aim of AgentDesktop

NA

December 29, 2025

1 Introduction

A useful way to frame the modern AI landscape is along two axes: how difficult a task is for humans and how difficult it is for machines. Notably, many capabilities that are effortless for people—perception, spatial reasoning, and embodied interaction—remain comparatively challenging for AI systems, even as AI excels in domains humans find difficult [?, ?]. This proposal focuses on one such gap: *structured visual understanding*, i.e., interpreting images that contain compositional, interactive structure and converting that understanding into reliable, executable actions. A primary instance of this challenge is computer vision over graphical user interfaces (GUIs), where the goal is not merely to "recognize objects," but to identify and relate UI elements (buttons, text fields, menus, icons), infer their functional roles (affordances), and ground language instructions or plans into precise on-screen actions (clicks, drags, keyboard input) [?, ?].

Recent advances in vision have been driven by architectures and generative models that substantially improve recognition and synthesis—including Vision Transformers (ViTs) [?] and diffusion models [?, ?]. However, GUI and structured-image understanding stress a different set of requirements: fine-grained spatial localization, layout reasoning, and action grounding. This gap is reflected in a growing body of benchmarks that explicitly evaluate *image-based navigation* and coordinate/action prediction. Early web-interaction environments such as MiniWoB/MiniWoB++ require agents to interact with rendered screens via mouse and keyboard actions [?, ?]. More recent work continues this pixel-to-action paradigm, including models trained to follow GUI instructions where click locations are represented explicitly (e.g., discretized coordinate bins) [?]. In parallel, GUI grounding benchmarks such as ScreenSpot evaluate whether a model can locate the correct target element in a screenshot given a natural language instruction [?]. Moving to more realistic and high-resolution professional settings, ScreenSpot-Pro reports that existing GUI grounding approaches still perform poorly, with the best model achieving only 18.9% on its benchmark [?]. Desktop-centric evaluations further highlight the same limitations: UI-Vision provides dense annotations across 83 real-world desktop applications (bounding boxes, labels, and action trajectories such as clicks, drags, and typing) and defines tasks for Element Grounding, Layout Grounding, and Action Prediction; its evaluations expose persistent failures in spatial reasoning and complex interactions such as drag-and-drop [?]. Complementary agent benchmarks (e.g., WebArena, Mind2Web, and Windows Agent Arena) reinforce that reliable end-to-end computer use requires robust screen understanding and grounded interaction, not just recognition [?, ?, ?].

These limitations restrict AI systems from performing many high-value tasks that inherently involve interacting with software: automated testing, user assistance, and robotic process automation. We argue that building computer vision models capable of interpreting structured data in images—and grounding that understanding into reliable, low-level UI actions—can unlock entire domains of automation.

For example, within software development life-cycle workflows, AI systems could operate in a sandboxed environment to extensively test applications by navigating forms, clicking buttons, and verifying outputs. In user assistance scenarios, AI could help users navigate complex software by understanding UI layout and providing context-aware guidance, or even performing tasks on a user's behalf. More broadly, improving structured visual understanding also supports native interpretation of structured scientific visuals (e.g., charts and plots embedded in technical documents), enabling AI systems to extract and act on information presented in non-textual formats. In other words, by addressing the challenge of structured visual understanding, we can significantly enhance the practical capabilities of AI systems.

2 References

References

[1] Hans Moravec. *Mind Children: The Future of Robot and Human Intelligence*. Harvard University Press, 1988.

[2] Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253, 2017. `https://doi.org/10.1017/S0140525X16001837`.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. arXiv:2010.11929. `https://arxiv.org/abs/2010.11929`.

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. arXiv:2006.11239. `https://arxiv.org/abs/2006.11239`.

[5] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. 2022. arXiv:2204.06125. `https://arxiv.org/abs/2204.06125`.

[6] Tianlin Tim Shi, Andrej Karpathy, Lin Xiao, and Percy Liang. World of Bits: An open-domain platform for web-based agents. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017. `https://proceedings.mlr.press/v70/shi17a/shi17a.pdf`.

[7] Evan Zheran Liu, Kelvin Guu, Panupong Pasupat, Tianlin Shi, and Percy Liang. Reinforcement learning on web interfaces using workflow-guided exploration. In *International Conference on Learning Representations (ICLR)*, 2018. arXiv:1802.08802. `https://arxiv.org/abs/1802.08802`.

[8] Peter Shaw, Ming-Wei Chang, and Kenton Lee. Learning to follow instructions via graphical user interfaces. 2023. arXiv:2306.00245. https://arxiv.org/abs/2306.00245.

[9] 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 (ACL)*, 2024. arXiv:2401.10935. https://arxiv.org/abs/2401.10935.

[10] Kaixin Li *et al.* ScreenSpot-Pro: GUI grounding for professional high-resolution computer use. 2025. arXiv:2504.07981. https://arxiv.org/abs/2504.07981.

[11] Shravan Nayak, Xiangru Jian, Kevin Qinghong Lin, Juan A. Rodriguez, Montek Kalsi, Rabiul Awal, Nicolas Chapados, M. Tamer Özsu, Aishwarya Agrawal, David Vazquez, Christopher Pal, Perouz Taslakian, Spandana Gella, and Sai Rajeswar. UI-Vision: A desktop-centric GUI benchmark for visual perception and interaction. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. arXiv:2503.15661. https://arxiv.org/abs/2503.15661.

[12] Shuai Zhou *et al.* WebArena: A realistic web environment for building autonomous agents. 2023. arXiv:2307.13854. https://arxiv.org/abs/2307.13854.

[13] Xiang Deng *et al.* Mind2Web: Towards a generalist agent for the web. 2023. arXiv:2306.06070. https://arxiv.org/abs/2306.06070.

[14] Rogerio Bonatti *et al.* Windows Agent Arena: Evaluating multi-modal OS agents at scale. 2024. arXiv:2409.08264. https://arxiv.org/abs/2409.08264.