EMBODIED WEB AGENTS: Bridging Physical-Digital Realms for Integrated Agent Intelligence

Yining Hong* **Rui Sun*** **Bingxuan Li†** **Xingcheng Yao†** **Maxine Wu†** **Alexander Chien†**
Da Yin **Ying Nian Wu** **Zhecan James Wang** **Kai-Wei Chang**

University of California, Los Angeles

Abstract

AI agents today are mostly siloed — they either retrieve and reason over vast amount of digital information and knowledge obtained online; or interact with the physical world through embodied perception, planning and action — but rarely both. This separation limits their ability to solve tasks that require integrated physical and digital intelligence, such as cooking from online recipes, navigating with dynamic map data, or interpreting real-world landmarks using web knowledge. We introduce EMBODIED WEB AGENTS, a novel paradigm for AI agents that fluidly bridge embodiment and web-scale reasoning. To operationalize this concept, we first develop the EMBODIED WEB AGENTS task environments, a unified simulation platform that tightly integrates realistic 3D indoor and outdoor environments with functional web interfaces. Building upon this platform, we construct and release the EMBODIED WEB AGENTS Benchmark, which encompasses a diverse suite of tasks including cooking, navigation, shopping, tourism, and geolocation — all requiring coordinated reasoning across physical and digital realms for systematic assessment of cross-domain intelligence. Experimental results reveal significant performance gaps between state-of-the-art AI systems and human capabilities, establishing both challenges and opportunities at the intersection of embodied cognition and web-scale knowledge access. All datasets, codes and websites are publicly available at our project page https://embodied-web-agent.github.io/.

1 Introduction

Recently, we have seen the proliferation of web agents capable of retrieving information online [Shi et al., 2017, Yao et al., 2022, Deng et al., 2023, Zhou et al., 2023, Koh et al., 2024] — yet they remain confined to screens disembodied from the real world. Meanwhile, their physical counterparts — robots and embodied systems — navigate the world but with limited access to the Internet. What if the boundary between the digital and physical realms were shattered? What if web agents stepped out of the browser, with keys to perceive and act in the real 3D physical world, while physical robots autonomously tapped into the encyclopedic knowledge of the web? As illustrated in Figure 1, such agents would not only assess the ingredients in your kitchen, search for matching recipes online, shop for missing items, and cook your favorite dish for you; but also traverse historical landmarks, interpret architectural styles using both their own perception and Wikipedia, leave personalized reviews, and perhaps even return with a souvenir in hand. We, as humans, don't compartmentalize our intelligence into "physical-only" and "digital-only" modules — we fluidly move between realms. What if contemporary AI agents could likewise achieve the best of both worlds?

Building such agents *goes far beyond a mere combination of isolated web and embodied systems*; it presents a set of deeply intertwined challenges. The first is *the perceptual grounding problem*: how can an agent link abstract digital instructions (e.g., "cook potato and egg until golden brown" as in

Preprint. Under review.

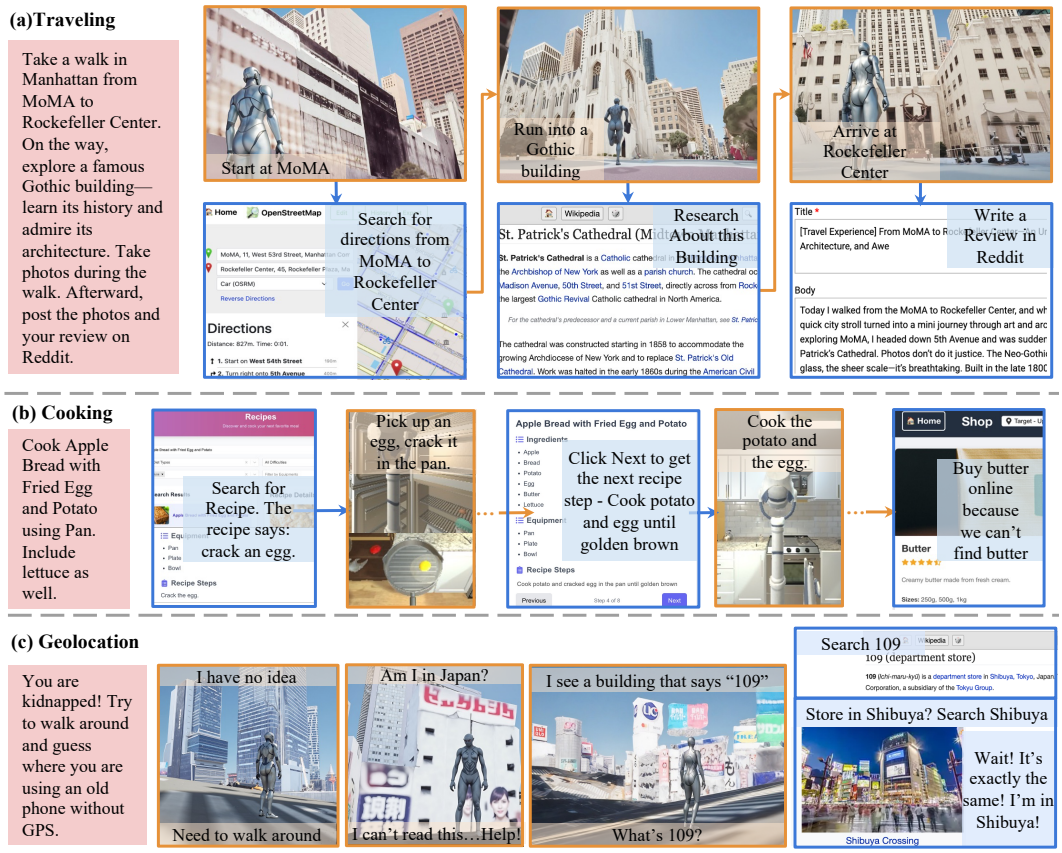

Figure 1: **Illustrative examples of our EMBODIED WEB AGENTS conceptual paradigm, tasks and environments.** Blue boxes and arrows indicate web interaction / switching to the web respectively. Orange boxes and arrows indicates acting in / switching to the embodied environment. We omit most intermediate actions due to the large number of interaction steps.

Figure 1 (b)) with the high-dimensional data streams of the physical world (e.g., visually recognizing the transition of potatoes and eggs to a golden state through a series of embodied observations)? Addressing this requires embodied perception, where agents actively interpret their surroundings through movement, interaction, and multimodal sensing — continually acquiring feedback from their environment and aligning these observations with digital instructions. The second challenge is *cross-domain planning*: how should an agent decide when to shift between physical actions and digital information retrieval, particularly when information from one domain contradicts or supplements the other? For instance, the online map may suggest a path to visit Rockefeller Center, but real-world observation may reveal that the center is closed due to a protest, demanding a dynamic reevaluation of the agent's plan. To navigate seamlessly between domains, agents must maintain a coherent and persistent representation that bridges physical and digital contexts — recalling physical experiences when operating online, and retrieving digital knowledge when acting in the world. Despite all these challenges, there remains a surprising lack of research targeting this level of integrated intelligence — both in terms of conceptual frameworks and benchmark development. As a result, progress in each domain often unfolds in isolation, with limited cross-pollination between the two paradigms.

To this end, we introduce EMBODIED WEB AGENTS as a new conceptual paradigm of AI systems that unify physical embodiment with web-scale knowledge access — capable of perceiving and acting in the real world while reasoning over dynamic, unstructured information from the web. To operationalize this concept, we first develop the EMBODIED WEB AGENTS task environments, a unified simulation platform that integrates realistic 3D environments with interactive web interfaces. This platform combines (1) indoor settings from AI2-THOR, (2) outdoor navigation in Google Earth, and (3) web interfaces including Wikipedia, online stores, recipe websites, map services *etc.*, enabling agents to interact seamlessly with both physical and digital spaces. Building upon this environment,