

Accessible Human-AI Collaboration

Mina Huh

minahuh@berkeley.edu

Electrical Engineering and Computer Sciences

University of California, Berkeley

Berkeley, CA, USA

My Dissertation: Accessible Human-AI collaboration across diverse perception and behavior

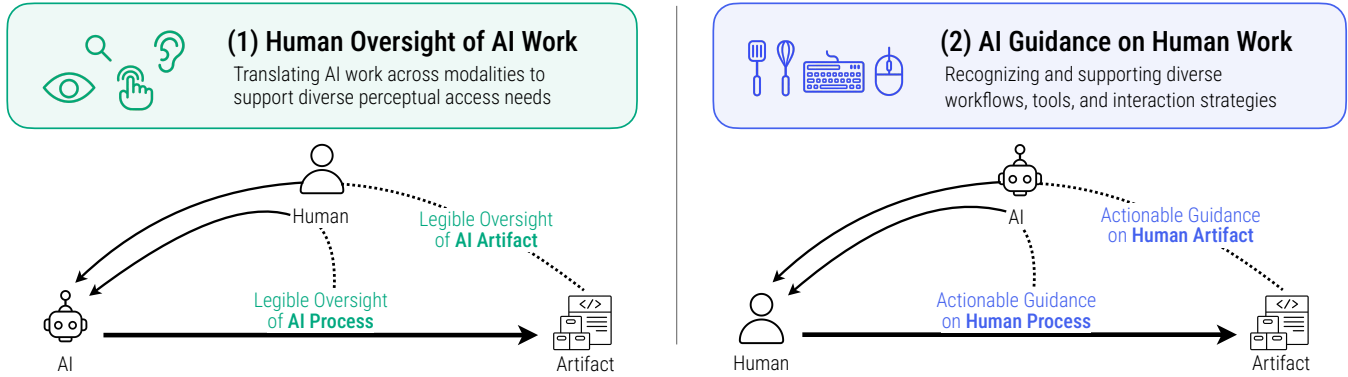


Figure 1: My dissertation explores accessible Human-AI collaboration across diverse perception and behavior. (1) For diverse perception, I make AI work legible by translating it across modalities, supporting oversight of AI-generated artifacts [4, 6, 9] and AI processes [5, 7]. (2) For diverse behavior, I make AI guidance actionable by grounding it in accessibility data of users' workflows, tools, and strategies, supporting guidance on human-created artifacts [8, 12] and human work processes [2, 11].

Abstract

Humans and AI increasingly collaborate to complete everyday tasks, learn new skills, and co-create artifacts. For effective collaboration, users must understand AI work in order to verify and intervene, while AI systems must understand human work in order to provide relevant guidance. Yet, current AI collaborators often rely on visual or text-heavy communication channels that presume particular perceptual abilities, making AI work difficult to perceive and interpret for users with disabilities (e.g., blind users). AI collaborators also assume normative interaction patterns, making their guidance less applicable to users who work through different tools, strategies, or assistive technologies. My dissertation addresses this accessibility gap by designing Human-AI collaboration across diverse perception and behavior. I explore two complementary thrusts: (1) legible oversight of AI work and (2) actionable AI guidance for human work. My research support people not only to access AI systems, but to effectively collaborate with them.

ACM Reference Format:

Mina Huh. 2026. Accessible Human-AI Collaboration. In *The 39th Annual ACM Symposium on User Interface Software and Technology (UIST Adjunct '26)*, November 02–05, 2026, Detroit, MI, USA. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3830397.3841705>

'26), November 02–05, 2026, Detroit, MI, USA. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3830397.3841705>

1 Introduction

AI systems are increasingly becoming *collaborators* in our everyday work. Beyond answering questions or automating isolated commands, AI can now guide people through tasks [2, 10, 11], review and revise artifacts [8, 12], generate creative outputs [4, 6, 9], and execute multi-step actions on users' behalf [7, 14, 20]. Through collaboration, human and AI complement each other: people bring goals, lived experience, contextual judgment, and values, while AI can operate at speed, retain and retrieve large amounts of information, and carry out repetitive, complex steps at scale. While broadly useful, AI collaborators create important opportunities for people with disabilities. Compared to manual assistance, AI support can be more readily available, scalable, and personalized to users' access needs [3, 18, 19]. Yet today's AI collaborators are not inherently accessible, limiting who can fully benefit from collaboration.

Effective collaboration requires both sides to read and respond to the evolving state of work. Yet, current AI systems rely on two assumptions that limit accessible Human-AI collaboration. First, they assume **uniform perceptual access**: that users can follow AI process and outputs through standard visual, text-heavy, or fast-moving representations, despite differences in perceptual and attentional capacity [7, 11]. As a result, users struggle to verify, intervene, or steer the AI, reducing their agency. Second, they assume **normative interaction patterns**: that users act, use tools, and express intent in ways reflected in dominant training data and



This work is licensed under a Creative Commons Attribution 4.0 International License. *UIST Adjunct '26, Detroit, MI, USA*

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2855-6/2026/11

<https://doi.org/10.1145/3830397.3841705>

interface conventions [15, 16]. This can misrepresent disability-adapted strategies, assistive-tool use, or alternative environmental setups. As a result, users receive guidance that is not actionable or misapply AI suggestions, increasing effort and errors.

In my dissertation, I explore accessible Human–AI collaboration across diverse perception and behavior. I organize this work around two complementary directions (Figure 1). First, when AI systems work on behalf of users, I develop *legible oversight* mechanisms that help users perceive, verify, and steer AI work. Second, when users lead the work and AI provides support, I develop *actionable guidance* grounded in diverse workflows, tools, strategies, and access needs. Across my works, I pursue:

- **Legible oversight of AI work.**
 - **AI artifacts:** I build systems that help blind and deaf creators verify, compare, and refine AI-generated images [9], videos [6], and music [4].
 - **AI processes:** I co-design multimodal communication that helps blind and low-vision users monitor agent actions, progress, and failures across physical robots [7] and computer-use-agents [5].
- **Actionable guidance for human work.**
 - **Human artifacts:** I build systems that make visual feedback understandable and actionable for blind creators in video editing [12] and web design [8].
 - **Human processes:** I model diverse user strategies from accessibility-specific workflow data and build proactive coach systems for blind users in cooking [11] and dance [2].

Together, these directions show that accessibility in Human-AI collaboration requires access not only to AI tools, but to the collaborative process itself. My thesis shows how AI systems can remain effective, robust, and agency-preserving across diverse users.

2 Legible Human Oversight of AI Work

When AI systems work on users’ behalf, users need accessible ways to understand, verify, and steer what the system produces and how it acts. My work explores this through sensemaking support for reviewing AI-generated outputs and process communication for monitoring AI agents as they work.

Sensemaking of AI-Generated Outputs. Since the public rise of text-to-image models in 2022, generative AI has promised to make image creation more accessible: users can generate visual content simply by writing a prompt. However, for blind and low-vision creators, accessibility barriers shift from *creating* images to *reviewing* them. While prior work has focused on helping blind users understand images as consumers, reviewing images as creators requires understanding them in context to their creative intent. Creators need to verify whether generated images match their intent, identify visual errors or distortions, and compare multiple candidates before deciding how to refine a prompt or select the final output.

To address this, we presented *GenAssist* [9], a system that helps blind creators review generated images and compare alternatives (Figure 2). *GenAssist* proactively predicts visual questions blind

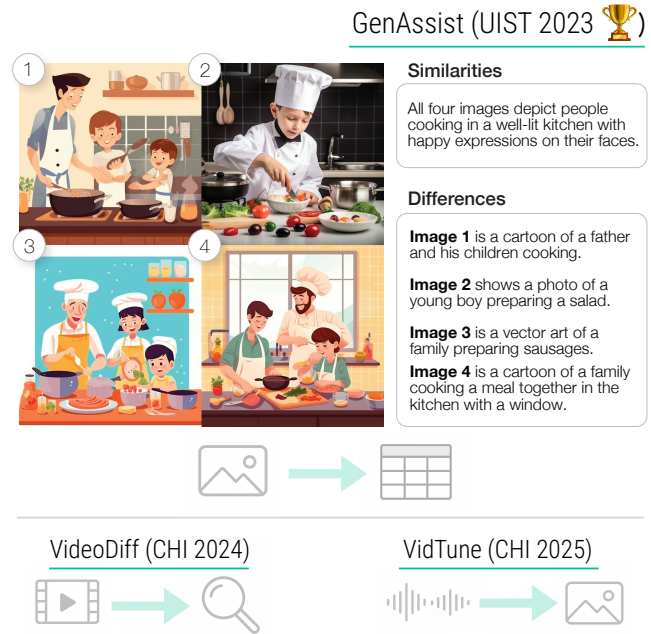


Figure 2: Systems that help blind creators verify and compare AI-generated images [9] and videos [6], and help deaf creators review generative music [4].

creators may ask about each image, answers them with vision-language models, and presents aggregated summaries in a screen-reader-friendly table for skimming and comparing candidates. Descriptions from *GenAssist* captured twice as many image differences as human-written descriptions. In a study with 12 blind creators, participants found *GenAssist* more useful than a baseline for comparing images, also detecting when outputs did not align with their intent, and steering the image-generation process.

GenAssist received a Best Paper Award at UIST 2023 and led to product impact through alt-text support for Adobe Firefly image generations [1]. In collaborations with Adobe Research, I extended its core idea of surfacing differences among generative alternatives through modality translation on other generative media. *VideoDiff* [6] supports low-vision creators in reviewing AI-edited videos, and *VidTune* [4] helps deaf and hard-of-hearing creators assess generative music through visual thumbnails.

Monitoring of AI Agent Processes. Beyond reviewing AI-generated outputs, accessible collaboration also requires understanding AI work as it unfolds. As AI systems increasingly act as agents that complete multi-step tasks on users’ behalf, accessibility depends not only on the final output, but also on whether users can monitor progress, detect errors, and intervene.

I first examine this challenge in robotics, where accessibility research has often focused on robots designed specifically to assist people with disabilities (e.g., navigation or feeding robots). Less explored is whether *general-purpose robots* that automate tasks for everyone are themselves accessible to blind and low-vision users. My recent work studied how blind users monitor such robots during everyday tasks [7]. In an observational study with 10 blind

Robot Oversight (CHI 2026 InterAI 🏆)



Legible Agents (Ongoing)

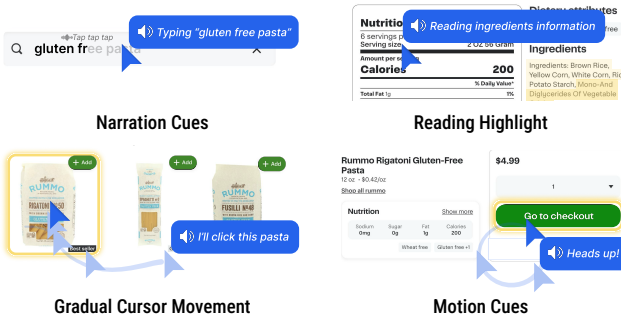


Figure 3: Accessible oversight across embodied and software agents: (Top) Blind users deploy technological and sensory strategies to monitor robot actions and status [7]. (Bottom) We explore multimodal cues that make computer-use agent actions more legible [5].

participants using two robot platforms, participants relied on non-visual cues (e.g., sound during execution, touch afterward, Figure 3). However, these cues were often ambiguous, leading to missed errors and uncertainty about task outcomes. Existing visual interpretation tools provided generic scene descriptions, but not task-specific information about robot progress or errors. As a result, participants reported an average of 7.5 inaccuracies in understanding robot behavior ($SD=1.27$), including unnoticed errors (51%) (e.g., when the robot bringing the wrong object) and incorrect assumptions (30%) e.g., assuming the robot had closed a bottle cap). To address this, I am building a multimodal robot communication pipeline that preserves blind users’ existing verification strategies while making ambiguous robot behavior clearer. The system proactively communicates task progress and failures, while allowing users to pause and verify robot state through touch, sound, or questions.

Beyond robots, computer-use-agents can also autonomously carry out multi-step tasks on users’ behalf. Yet unlike robots whose physical actions can be tracked through sound or touch, computer-use agents provide few non-visual cues for understanding what they are doing. In a formative analysis of 18 agent interfaces, we found that agents were often abrupt with under-cued screen actions,

and that explanations were buried in long text logs. To make these processes more legible, I am developing legible agents [5] for users with diverse access needs (Figure 3). It uses multimodal cues that communicate agent actions as they unfold, including action narration, sound cues, gradual cursor movement, target highlighting, and motion cues before high-stakes actions.

3 Actionable AI Guidance for Human Work

While AI becomes increasingly capable automating work, it is equally important to support people in doing the work themselves. Users may want to retain creative agency [8, 12], build skills [11], and stay engaged in the process [2]. Yet, accessibility barriers can make it difficult to understand how to improve or what to do next. I explore how AI can provide actionable support through two forms of guidance: feedback that helps users improve their artifacts, and situated guidance that helps users make progress.

Feedback on Human-Created Artifacts. When users create visual artifacts themselves, they need ways to review their work quality, identify issues, and decide how to revise. In *AVscript* [12], I supported accessible video editing for blind creators who lacked efficient access to a video’s visual content, quality, and structure. Based on interviews with 8 blind video creators, we developed an *audio-visual script* that embeds both speech and visual information into a text-based representation that can be easily navigated with a screen reader. *AVscript* identifies scenes, objects, and visual errors (e.g., blur or poor lighting), and places them as editable units within the script (Figure 4). This lets blind creators navigate through visual footage using familiar text navigation, identify relevant moments and visual issues, and make their own editing decisions. In a study with 12 blind creators, participants reported 52.6% less effort, 55.7% more confidence in their final videos, and an 81.8% increase in independence when reviewing their work compared with their own tools. The system was later used in a film class workshop at the Texas School for the Blind and Visually Impaired.

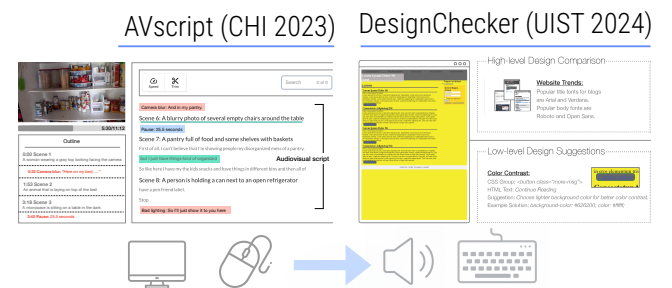


Figure 4: Systems that make visual feedback understandable and actionable with screen readers for blind video creators [12] and web developers [8].

I extended this work to web design, where developers must also understand how low-level code shape the overall look and feel of a website and whether it aligns with visual conventions. I explored this through *DesignChecker* [8], a browser extension that helps blind developers evaluate and improve the visual design of their websites. Our analysis of 20 websites created by blind developers,

together with interviews with 9 blind developers, showed that they wanted support assessing visual usability for sighted users and alignment with common design standards. It was difficult to use screen readers to detect issues such as illegible text, misaligned elements, and inharmonious colors (e.g., a screen reader can reveal that text is orange on a yellow background, but not whether that combination is illegible). DesignChecker provides actionable CSS suggestions and connects each suggestion to relevant visual design guidelines and examples from reference websites. This helps blind developers both address individual issues and build visual design knowledge. In a study with 8 blind developers, participants using DesignChecker made 82% more edits and resolved 141% more design issues than with their existing tools. Together, these systems show how visual feedback can be understandable and actionable, enabling blind creators to make creative decisions rather than delegating them to AI or relying on sighted collaborators.

Vid2Coach (UIST 2025)



Figure 5: Vid2Coach [11] transforms tutorial videos into a smart-glasses assistant that provides instructions, progress-aware feedback, and answers to user questions.

Guidance During Human Task Processes. Beyond reviewing artifacts, users also need guidance while carrying out tasks. How-to videos are a common source of such guidance, but rely on visual demonstrations that are inaccessible to blind users (e.g., “Turn off the heat when the butter turns brown”). Even when told a user is blind, off-the-shelf AI models often suggest asking a sighted person for help rather than offering non-visual strategies.

We first observed how vision rehabilitation therapists guide blind people through video-based tasks. They combined detailed descriptions, non-visual workarounds, and progress feedback. Based on these insights, Vid2Coach [11] transforms how-to videos into wearable, camera-based task assistants on smart glasses (Figure 5). Vid2Coach first extracts high-level task steps from a tutorial video, then enriches each step with demonstration details and completion criteria using multimodal understanding. To provide accessible workarounds personalized to users, it uses retrieval-augmented generation over 280 curated accessibility resources on common physical tasks (e.g., cooking, crafting). As users perform the task, Vid2Coach monitors egocentric stream to provide real-time, mixed-initiative feedback. In an at-home study with 8 blind participants,

Vid2Coach helped them complete cooking tasks with 58.5% fewer errors than their typical workflow.

I further extended this direction by mentoring an undergraduate student on a project exploring accessible learning for more expressive physical activities, specifically dance [2]. In a co-design study with 28 blind learners, dance teachers, audio describers, and multimodal interaction designers, we explored how dance instruction can move beyond narrations toward multimodal guidance that conveys spatial relations and expressive qualities. Together, these projects show how AI guidance can be grounded in non-visual expertise rather than simply translating visual instructions.

4 Ongoing and Future Work

My ongoing work extends the dissertation by making legibility 1) measurable and 2) intrinsic to autonomous agents. First, I am developing methods to *measure and predict* the legibility of computer-use agents. My prior work shows that these agents can be difficult to monitor [5], but there is currently no systematic way to evaluate how understandable their behavior is to users. Computer-use agents can complete the same task in very different ways [13]. Some rely on visible GUI actions, while others use subprocesses or shortcuts that are largely invisible to users. I am developing benchmarks and human-aligned metrics that measure how these choices affect users’ ability to understand, predict, and intervene in agent behavior. This can support agents that optimize not only for task success, but also for legibility to the users overseeing them.

Second, I am exploring how accessibility can become part of *agent planning* itself. Current robot communication is often generated after the robot has decided how to complete a task (e.g., with a camera and VLM that describes robot behaviors [7, 17]), explaining a fixed sequence of actions. However, for accessible monitoring, the robot may need to change how it acts (e.g., pausing at a verifiable state, signaling uncertainty before acting, choosing an action that is easier for the user to perceive). I am developing communication-aware planning methods that jointly generate physical actions and communication primitives as part of the robot’s executable plan. This allows robots to plan for both task success and whether users can oversee and intervene in their behavior. Together, these directions move accessibility from adapting agent behavior after the fact to shaping how autonomous agents act and communicate.

Looking ahead, I will extend my framework for accessible Human-AI collaboration to longer-term and team-based settings. I will study how legible oversight and actionable guidance evolve as users and AI systems adapt to one another, and how these mechanisms change in mixed-ability teams where blind users, sighted collaborators, and AI systems all work together.

5 Conclusion

My thesis demonstrates how Human-AI collaboration can become more accessible through legible oversight of AI work and actionable guidance for human work. Across systems for reviewing AI-generated artifacts, monitoring agent behavior, and guiding users through creative and physical tasks, I show that designing for diverse perception and behavior can make collaboration more effective, robust, and agency-preserving. Overall, my research re-frames accessibility as active participation in Human-AI collaboration rather than passive access to AI assistance.

Acknowledgments

I am grateful to my advisor, Amy Pavel, for supporting me throughout the course of the Ph.D. I also appreciate many mentors, collaborators, and friends for their wisdom and encouragement. This research is supported by Google Ph.D. Fellowship and Adobe Research.

References

- [1] Adobe. [n. d.]. Adobe Firefly — Free Generative AI for creatives. <https://www.adobe.com/products/firefly.html>. Accessed Dec 16, 2025.
- [2] Ujjaini Das, Shreya Kappala, Meng Chen, Mina Huh, and Amy Pavel. 2026. Co-Designing Multimodal Systems for Accessible Asynchronous Dance Instruction. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–18.
- [3] Krzysztof Gajos and Daniel S Weld. 2004. SUPPLE: automatically generating user interfaces. In *Proceedings of the 9th international conference on Intelligent user interfaces*. 93–100.
- [4] Mina Huh, C Ailie Fraser, Dingzeyu Li, Mira Dontcheva, and Bryan Wang. 2026. VidTune: Creating Video Soundtracks with Generative Music and Video-Based Thumbnails. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–25.
- [5] Mina Huh, Suk Min Hwang, and Amy Pavel. [n. d.]. Legible Agents: A Design Space for Overseeing Computer Use Agents. In *COLM 2026 The 2nd Workshop on Lifelong Agents: Learning, Aligning, and Evolving*.
- [6] Mina Huh, Ding Li, Kim Pimmel, Hujung Valentina Shin, Amy Pavel, and Mira Dontcheva. 2025. Videodiff: Human-ai video co-creation with alternatives. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–19.
- [7] Mina Huh, Huihan Liu, Albert Yu, Roberto Martín-Martín, Yuke Zhu, Maya Cakmak, and Amy Pavel. 2026. Designing Accessible Robot Communication for Blind People. In *The 3rd InterAI Workshop: Interactive AI for Human-centered Robotics*.
- [8] Mina Huh and Amy Pavel. 2024. DesignChecker: Visual design support for blind and low vision web developers. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–19.
- [9] Mina Huh, Yi-Hao Peng, and Amy Pavel. 2023. GenAssist: Making Image Generation Accessible. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) (UIST '23). Association for Computing Machinery, New York, NY, USA, Article 38, 17 pages. doi:10.1145/3586183.3606735
- [10] Mina Huh, Fangyuan Xu, Yi-Hao Peng, Chongyan Chen, Hansika Murugu, Danna Gurari, Eunsol Choi, and Amy Pavel. 2024. Long-Form Answers to Visual Questions from Blind and Low Vision People. In *First Conference on Language Modeling*.
- [11] Mina Huh, Zihui Xue, Ujjaini Das, Kumar Ashutosh, Kristen Grauman, and Amy Pavel. 2025. Vid2Coach: Transforming How-To Videos into Task Assistants. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*. 1–24.
- [12] Mina Huh, Saellyne Yang, Yi-Hao Peng, Xiang'Anthony' Chen, Young-Ho Kim, and Amy Pavel. 2023. AVscript: Accessible Video Editing with Audio-Visual Scripts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.
- [13] Hongrui Jia, Jitong Liao, Xi Zhang, Haiyang Xu, Tianbao Xie, Chaoya Jiang, Ming Yan, Si Liu, Wei Ye, and Fei Huang. 2026. Osworld-mcp: Benchmarking mcp tool invocation in computer-use agents. In *International Conference on Learning Representations*, Vol. 2026. 63734–63768.
- [14] Geunwoo Kim, Pierre Baldi, and Stephen McAleer. 2023. Language models can solve computer tasks. *Advances in Neural Information Processing Systems* 36 (2023), 39648–39677.
- [15] Kelly Avery Mack, Rida Qadri, Remi Denton, Shaun K Kane, and Cynthia L Bennett. 2024. "They only care to show us the wheelchair": disability representation in text-to-image AI models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–23.
- [16] Daniela Massiceti, Camilla Longden, Agnieszka Slowik, Samuel Wills, Martin Grayson, and Cecily Morrison. 2024. Explaining CLIP's performance disparities on data from blind/low vision users. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12172–12182.
- [17] Zihan Wang, Brian Liang, Varad Dhat, Zander Brumbaugh, Nick Walker, Ranjay Krishna, and Maya Cakmak. 2024. I can tell what i am doing: Toward real-world natural language grounding of robot experiences. *arXiv preprint arXiv:2411.12960* (2024).
- [18] Linda Yilin Wen, Cecily Morrison, Martin Grayson, Rita Faia Marques, Daniela Massiceti, Camilla Longden, and Edward Cutrell. 2024. Find my things: personalized accessibility through Teachable AI for people who are blind or low vision. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–6.
- [19] Jacob O Wobbrock, Shaun K Kane, Krzysztof Z Gajos, Susumu Harada, and Jon Froehlich. 2011. Ability-based design: Concept, principles and examples. *ACM Transactions on Accessible Computing (TACCESS)* 3, 3 (2011), 1–27.
- [20] Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh J Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. 2024. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems* 37 (2024), 52040–52094.