Sidekick: Designing Communication for Effective Multitasking with Computer Use Agents

Ruei-Che Chang
UMich
Ann Arbor, MI, USA
rueiche@umich.edu

Wenqian Xu
UMich
Ann Arbor, MI, USA
wxtu@umich.edu

Dingzeyu Li
Adobe Research
Seattle, WA, USA
dinli@adobe.com

Bryan Wang
Adobe Research
Seattle, WA, USA
bryanw@adobe.com

Anhong Guo
UMich
Ann Arbor, MI, USA
anhong@umich.edu

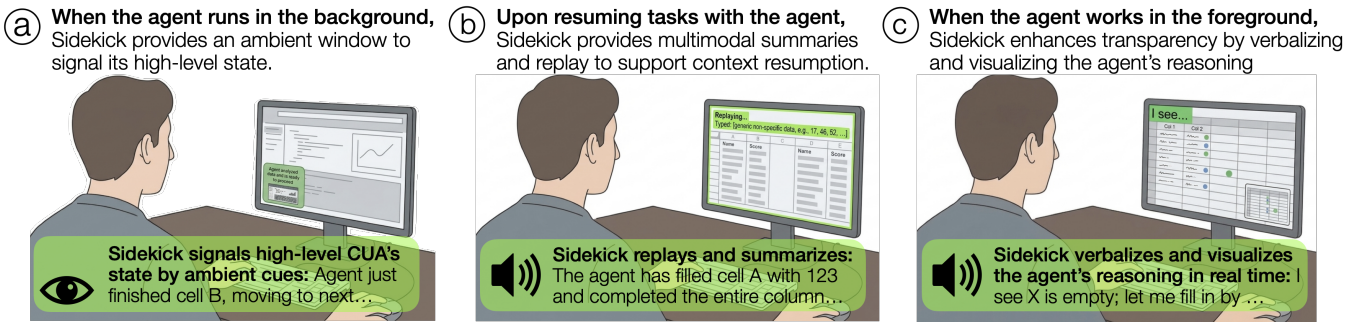

Figure 1: Sidekick supports effective multitasking with computer-use agents (CUAs) by bridging communication gaps across different stages of interaction. In this example, CUAs assist with data annotation for a paper the user is writing. (a) When *CUAs run in the background*, Sidekick signals their execution states through ambient cues, mitigating disruption to the user's primary task. (b) When *the user resumes interaction*, Sidekick presents multimodal summaries of completed actions, enabling rapid context resumption. (c) When *CUAs operate in the foreground*, Sidekick improves their transparency by verbalizing and visualizing the agent's reasoning process.

Abstract

Computer Use Agents (CUAs) can autonomously execute complex, multi-step tasks within GUIs, enhancing efficiency through parallel multitasking. However, our formative studies with CUA experts and GenAI users indicated that current feedback is primarily text-based, requiring sustained attention to monitor progress and offering limited visibility to trace past GUI interactions. Based on the findings, we developed a prototype system, Sidekick, for communicating CUAs' status with multimodal feedback across different stages of interaction: *(i)* When **CUAs run in the background**, Sidekick signals its execution state through ambient cues. *(ii)* **Upon resuming interaction with CUAs**, Sidekick provides multimodal summaries of completed actions to support rapid context resumption. *(iii)* When **CUAs operate in the foreground**, Sidekick enhances transparency by verbalizing and visualizing the agent's reasoning. A study with 30 participants demonstrated that Sidekick significantly improved multitasking performance with CUAs compared to baseline systems that presented textual feedback either in a typical chat or in an ambient display. Sidekick supported progress awareness, and error and action traceability more effectively. Finally, we demonstrate the promise of Sidekick through several example applications, and discuss implications for long-horizon human-agent collaboration.

UIST '26, Detroit, MI, USA

ACM ISBN 979-8-4007-2856-3/2026/11
https://doi.org/10.1145/3830398.3830476

CCS Concepts

• **Human-centered computing** → **Graphical user interfaces**; **User interface management systems**; **Auditory feedback**; **Information visualization**.

Keywords

Computer use agent, multimodal feedback, ambient display, LLM, VLM, multitasking, human-agent collaboration

ACM Reference Format:
Ruei-Che Chang, Wenqian Xu, Dingzeyu Li, Bryan Wang, and Anhong Guo. 2026. Sidekick: Designing Communication for Effective Multitasking with Computer Use Agents. In *The 39th Annual ACM Symposium on User Interface Software and Technology (UIST '26), November 02–05, 2026, Detroit, MI, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3830398.3830476

1 Introduction

Computer Use Agents (CUAs) operate through sequences of actions within graphical user interfaces (GUIs), making intermediate decisions, interacting across applications, and executing tasks on users' behalf. These capabilities hold significant promise for autonomously managing complex tasks [79, 94], where users can delegate and coordinate tasks according to their preferences. However, CUAs may operate for extended periods, during which users often switch to other tasks. In certain situations, such as those involving privacy-sensitive or high-stake decisions [37, 75, 79], CUAs may still require timely user input. As a result, users must manage multiple tasks concurrently while maintaining effective communication and oversight of CUAs to ensure successful task execution. This raises a

key question: *How should feedback be designed to support effective communication during multitasking with CUAs?*

To address this question, we conducted a formative study with three CUA experts and fifteen GenAI tool users (e.g., text, image, code generation). This allowed us to examine how issues observed in existing GenAI tools, which similarly involved AI reasoning, limited transparency, and extensive outputs, may transfer to CUAs, while grounding our analysis in CUA expert insights. We found that CUAs are often used in two interaction modes: **foreground**, where users actively monitor the agent's progress, and **background**, where agents run autonomously while users attend to other tasks. Across both modes, we identified several communication gaps.

When CUAs operate **in the foreground**, unlike other GenAI tools that display ongoing reasoning or real-time edits (e.g., vibe-coding tools), they provide limited transparency into their internal decision-making process and little GUI-visible feedback to support user engagement [37]. When running **in the background**, CUAs also lacked clear signals of progress or completion. As a result, they often remain detached from users' workflows and are checked asynchronously long after they are stuck or done. **Upon resuming interaction with CUAs**, users encountered feedback that remains largely text-based, similar to other GenAI systems. This made it difficult to reconstruct CUAs' progress, intervene in time, or understand how GUI states have evolved, particularly when CUAs perform consequential actions and produce artifacts beyond text.

To bridge this communication gap in multitasking with CUAs, we introduce Sidekick, a prototype system that supports monitoring and communicating CUA status across interaction stages: *(i)* When CUAs operate **in the background**, Sidekick conveys execution state with a high-level summary and visual thumbnail, along with ambient color cues in a peripheral display (e.g., green for normal progress, yellow for potential issues, red for being stuck). *(ii)* **Upon resuming interaction with CUAs**, Sidekick provides multimodal summaries that combine speech and visual highlights of key actions and system changes to support rapid context recovery. *(iii)* When CUAs operate **in the foreground**, Sidekick enhances transparency by verbalizing the agent's reasoning in real time and visualizing its actions (e.g., annotated clicks, typing, and screenshots).

To evaluate Sidekick, we conducted a mixed-methods study with 30 participants performing arithmetic questions as the primary task and spreadsheet filling as the secondary task. We found that Sidekick significantly improved multitasking performance compared to users working alone, and two baseline systems that presented textual feedback either in a typical chat or on a peripheral display. This improvement was driven by better outcomes in the CUA-assisted secondary task, where Sidekick enhanced users' awareness of the CUA's states and enabled timely intervention when errors occurred without disrupting users' primary tasks. Moreover, Sidekick reduced errors more effectively than the baseline systems by providing multimodal summaries of CUAs' completed actions upon users' return, which were perceived as more helpful for understanding and multitasking with CUAs. Finally, we demonstrated the promise of Sidekick through three application scenarios and discussed implications for long-horizon human-agent collaboration.

In summary, our work makes the following contributions:

(i) A formative study with 15 GenAI users and 3 CUA experts that identifies key communication gaps between humans and CUAs during multitasking.
(ii) Sidekick, a prototype system that bridges these gaps by providing feedback to support transparency, progress awareness, and context resumption in human-CUA multitasking.
(iii) A mixed-methods evaluation with 30 participants showing that Sidekick significantly improved multitasking performance with CUAs compared to chat-based or peripheral-text feedback without disrupting the primary task.
(iv) Design implications and three application scenarios to illustrate how Sidekick may be extended to future long-horizon human-agent collaboration.

2 Related Work

Our work builds on prior research in *(i)* computer-use and web agents that operate GUIs, *(ii)* transparency in human-agent collaboration, and *(iii)* ambient and peripheral displays that support multitasking, techniques for rapid context resumption.

2.1 Computer-Use and Web Agents

Recent advances in computer-use models have enabled agents that execute user goals by directly interacting with GUIs, bridging natural language intent and low-level actions (e.g., clicking, typing). Early work explored grounding models in web environments using screenshots, DOM structures, and action histories. For example, WebArena [96] and Mind2Web [46] introduced realistic web environments and benchmarks that map high-level prompts to executable interactions in standardized settings. Building on these, WebVoyager [52] and SeeAct [95] leveraged large multimodal models (LMMs) to support more open-ended web tasks.

Beyond web agents, "computer-use" (OS-level) agents generalize these capabilities to desktop environments. Given a user goal, they iteratively infer executable actions from screenshots in a closed loop grounded in GUI states. Industry systems (e.g., OpenAI Operator [12], Anthropic Claude [8], Google Gemini [13]) further demonstrate increasingly human-like interaction. Correspondingly, evaluation has shifted from static, outcome-based metrics to *human-centric*, trajectory-aware criteria that assess whether actions are visually grounded and contextually appropriate, which better captured real-world ambiguity and long-horizon tasks [18, 24, 49, 93].

These advances suggest that CUAs are becoming increasingly autonomous, capable of handling complex tasks at scale [79, 94]. However, increased autonomy also introduces risks, including errors and guardrails triggered by high-stakes actions [5, 75, 79], which still require timely user intervention. This raises a key question for the HCI community: *How should feedback be designed to support effective communication during multitasking with CUAs?* To address this, we conducted a formative study to derive design goals for Sidekick to support human-CUA multitasking.

2.2 Human-Agent Collaboration

Principles for mixed-initiative systems emphasized coupling automation with direct manipulation, allowing users to invoke, override, or terminate automation [54]. Human-AI guidelines similarly highlighted transparency, user correction, and error recovery [21].

Prior work shows transparency can improve performance and trust calibration, depending on task demands and interface integration [81, 87, 88]. Accordingly, explanation techniques in different modalities, such as textual (what, why, consequences) [33, 35] and visual (e.g., feature highlights [77], interactive visualizations [20, 40, 44, 91, 92]), have been developed to support human-AI collaboration [53, 84]. However, these approaches are underexplored for CUAs, which operate over long horizons and may hallucinate, get stuck, or cause irreversible errors, which require timely communication for user intervention [37, 72, 75]. To address this gap, Sidekick extends prior work by verbalizing and visualizing CUA reasoning and actions to enhance transparency in multitasking. Our study findings further highlight the need for context-aware, customizable communication to support effective human-CUA collaboration.

2.3 Ambient Display, Multitasking, and Context Resumption

HCI research on multitasking examined how users coordinated attention and managed interruptions. Prior work showed that interruption costs vary with timing [19] and task complexity [71]. Several systems have been developed to mitigate disruption by inferring cognitive load and scheduling notifications opportunistically [23, 31, 36, 55, 57, 58], or by supporting resumption through structured activity histories and video replay [47, 50]. Another line of work explored awareness without undue distraction. Peripheral and ambient displays presented "at-a-glance" information in the periphery, enabling background monitoring while maintaining focus [76]. These systems provided interfaces spanning on-screen sidebars [28], multi-monitor setups [68, 69], projections [66], and physical artifacts with ambient cues [41, 45]. Prior studies also compared presentation strategies in ambient displays, including text vs. graphics [82], motion vs. static [67], information density [43], and modality [22, 26]. These insights informed our design of ambient displays to unobtrusively signal CUAs' states in Sidekick and the baseline systems we compared in the evaluation (e.g., text in either chat or peripheral). Sidekick builds on prior work in structured history and video replay to support resumption for CUAs that produce *evolving GUI states*, rather than static or predetermined interfaces.

3 Formative Study

In this study, we aimed to investigate how users interact with CUAs and GenAI tools across prompting, waiting, and result retrieval, and what feedback supports such collaboration. We conducted a survey with 15 GenAI tool users (F1–F15; mean age = 26.5, SD = 3.3). The survey covered participants' experiences with text, image, and code generation, with a focus on workflow integration, sustained engagement, and how users revisit and interpret generated outputs. We also conducted 30–60-minute contextual inquiries with three CUA experts (F16–F18; mean age = 29.3, SD = 3.2), during which they demonstrated their typical workflows and shared observations about CUA use. F16 used the CUA product developed by Vercept [16], which was later acquired by Anthropic [2], for background tasks such as exploring and summarizing new papers and generating weekly stock charts. F17, a postdoctoral researcher, studied CUAs in assistive contexts, while F18, a PhD student, examined their use on complex cloud-based platforms. Collectively, their work spanned a range of CUA models and application contexts.

This mixed-sample approach reflects the relatively early stage of CUA development and adoption. As of 2025, available CUAs remained limited [5, 8, 13] and achieved approximately 60% accuracy even on basic tasks [14, 93], constraining opportunities for broader real-world use. Our approach therefore allowed us to examine how challenges observed in established GenAI tools may extend to CUAs while grounding these insights in expert practice.

3.1 Results

3.1.1 Some users treated GenAI tools as peripheral, background components in multitasking workflows. Users integrated GenAI tools into broader workflows alongside productivity tools (e.g., image generation for slides in PowerPoint or Keynote, and text generation for writing in Overleaf). They often used multiple GenAI tools concurrently to support multitasking. As a result, GenAI tools were typically treated as peripheral or background components, running in browser tabs, split-screen layouts, or small floating windows rather than occupying primary focus. In contrast, code generation tools were more tightly embedded within development environments (e.g., Cursor [7], Claude Code [3]). Similarly, CUAs were frequently executed in the background or in the cloud for scheduled or long-running tasks to avoid interfering with users' local computer and primary work (e.g., F16 used Vercept [16] to generate weekly stock market plots in Google Sheets).

3.1.2 Some users shifted attention to manually monitor progress. When tasks were complex or time-consuming, some users treated GenAI tools as background processes and shifted to other tasks (N=12), such as during *"refactoring or making big changes"* (F3). Because intermediate progress signals were limited, participants periodically checked progress by using Alt+Tab (F11, F12), switching back *"from time to time"* (F2), or monitoring screens and terminal messages (F17). Although disruptive, these checks helped users maintain context and avoid the cost of resuming interrupted work, as F5 explained: *"I would constantly check the progress because I don't want to lose my flow."* These findings highlight the need for clearer, real-time indications of GenAI progress.

3.1.3 Some users sustained attention to GenAI tools. Some users remained attentive to GenAI tools when reviewing outputs or handling high-stakes tasks. F8 stayed on ChatGPT *"for important tasks"* to pause or revise prompts. Others monitored only the initial output before returning to their main work; for example, F13 verified Claude's to-do list before switching away. Participants also used secondary monitors (F10) or split-screen layouts (F13) to maintain peripheral awareness. Fast inline feedback in GenAI tools further reduced attention shifts, as F4 noted: *"The inline suggestions appear quite quickly, so that I can use them immediately without shifting."* However, feedback in current CUA tools was often overwhelming, whether presented as dense terminal text (F17, F18) or as a combination of screenshots, actions, and reasoning (F16). These findings highlight the need for timely, unobtrusive feedback that supports lightweight monitoring.

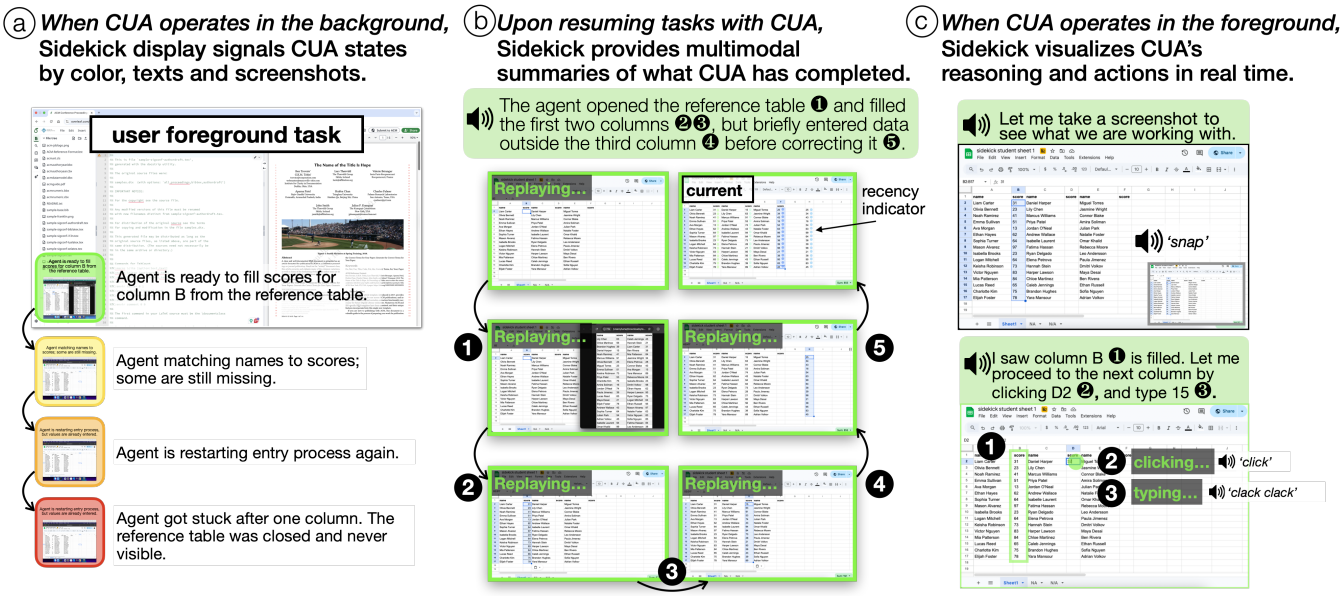

Figure 2: Sidekick's multimodal feedback across different stages of user interactions with CUAs. (a) When the CUA runs in the background, Sidekick provides peripheral awareness through ambient visual cues (colors, text, and thumbnails). Green indicates smooth progress, yellow signals minor errors, orange reflects accumulating issues, and red indicates the agent is stuck and has been paused for user intervention. (b) Upon resuming, Sidekick delivers multimodal summaries combining speech and visual replay, highlighting completed actions as well as encountered errors. (c) When the CUA is in the foreground, Sidekick visualizes the agent's reasoning and actions in real time, augmented with synchronized speech and foley sound effects.

3.1.4 Current interface features to support task switching are insufficient. To support task switching, particularly when returning to GenAI tools, participants relied on notifications to signal completion. However, prolonged waiting made re-engagement difficult, often requiring users to scroll through logs or chat histories to reconstruct context. As F10 noted, lengthy responses made it easy to *"lose track of where the conversation is going."* Interface structuring partially mitigated this issue. Elements such as *code blocks*, *bold headings*, and clear response semantics helped users reorient; for example, F3 described scanning high-level explanations before examining detailed edits. However, current CUA chat interfaces offer limited structured or layered feedback, making users rely primarily on final screen states or text-based reasoning. This highlights the need for better support for task resumption, especially as CUAs produce consequential actions beyond text.

4 Sidekick

Sidekick is a prototype system designed to bridge the communication gap across different interaction stages with CUAs. Based on the formative study findings, we derived three design goals:

- **D1** *When CUAs operate in the background*: Enable peripheral awareness of CUAs' intermediate progress to reduce manual checking (Sections 3.1.1 & 3.1.2).
- **D2** *Upon resuming tasks with CUAs*: Offer structured, high-level summaries of completed actions, system changes to support rapid context resumption (Sections 3.1.2 & 3.1.4).
- **D3** *When CUAs operate in the foreground*: Provide real-time feedback on CUAs' actions and reasoning to support transparency (Section 3.1.3).

Specifically, it provides peripheral signals to convey task progress when running in the background (Section 4.1), structured multimodal summaries to communicate completed actions and changes (Section 4.2), and real-time, synchronized text and visual feedback when CUAs operate in the foreground (Section 4.3). Below, we illustrate these Sidekick interactions across stages using an example in which a user writes a paper while a CUA fills a spreadsheet with relevant data (Figure 1).

4.1 CUAs operating in the background

To fulfill **D1** and support peripheral awareness of CUA progress, Sidekick introduces an adjustable, always-on-top peripheral window, referred to as the *Sidekick display*. The display communicates high-level CUA activity through concise text summaries, screenshots, and color-coded status signals (Figure 2a). Within this window, Sidekick condenses raw CUA messages into brief, one-sentence summaries (e.g., "Agent found data references, is analyzing the spreadsheet, and will fill Column B starting at B2"). These summaries allow users to monitor progress at a glance while remaining focused on their primary task. For later reference, the same message also appears in the chat interface (Figure 3d, shown in black). Sidekick evaluates each generated action by comparing screenshots taken before and after execution. A VLM (`gemini-2.5-flash`) assesses whether the resulting state matches the user's query or the CUA's current subgoal. An action is marked erroneous when the post-execution state fails to reflect the intended outcome. Most errors involved either *no progress*, where the interface remained unchanged, or an *incorrect action*, where the resulting state diverged

from the goal. Based on this definition, Sidekick tracks the number of consecutive errors, denoted by e, and adjusts the color of the *Sidekick display* to indicate system status: yellow when $e > x$, orange when $e > y$, and red when $e > z$. Upon reaching the red state, Sidekick automatically pauses the CUA as a safeguard. This design aims to draw the user's attention when intervention may be necessary while remaining minimally intrusive during normal operation. In our study, we set $x = 3$, $y = 6$, and $z = 8$ based on the error patterns observed during prototype development.

4.2 Resuming tasks with CUAs

To address **D2** and support context resumption, Sidekick generates a multimodal summary when users return to CUAs (e.g., focusing on the CUA's tasks), describing actions completed during background execution (Figure 2b). To achieve this, Sidekick operates continuously in the background and produces summaries at each completed step by prompting a VLM with prior actions, step indices, screenshots, and agent messages. The summaries are converted into the Speech Synthesis Markup Language (SSML) [15] with embedded step IDs, allowing text-to-speech (TTS) services to generate timestamps for synchronized playback. Using these, Sidekick aligns narration with corresponding screenshots and action visualizations (Detailed in Section 4.4). Depending on user preference, summaries can be presented as either audio or text. The CUA remains paused during summary playback, allowing users to review the completed actions before execution resumes. After playback, Sidekick visualizes action recency using a color gradient (e.g., green for earlier actions, blue for recent ones), helping users understand how prior steps led to the current system state.

4.3 CUAs operating in the foreground

To fulfill **D3** and provide transparent, real-time feedback, Sidekick externalizes the CUA's reasoning, plans, and actions through multimodal cues when users actively engage with the agent (Figure 2c). This includes verbal and visual feedback to accommodate different attention preferences and interaction contexts. Specifically, Sidekick verbalizes the CUA's messages, including internal planning (e.g., "I can see column B is empty and needs to be filled. Now I need to reference the Preview app to get the correct scores.") and action-completion updates (e.g., "I've typed 65 in cell D7. Now I'll press Enter to confirm the entry."). It also generates Foley sounds [9, 38] to accompany computer interactions such as clicking, typing, and taking screenshots. In parallel, Sidekick visualizes actions on the screen, in sync with the speech and foley sounds. For example, it highlights spoken content with cropped bounding boxes (Figure 2.1), annotates clicked locations with visual markers (Figure 2.2), and displays the issued keyboard commands in texts (Figure 2.3). Similar to multimodal summaries, this synchronization is achieved by converting messages into SSML [15] with embedded action labels, enabling TTS to generate timestamps that align speech with visualizations. We detail the implementation in the next section.

4.4 Implementation details

We use the existing CUA framework [6] and the model `anthropic/-claude-sonnet-4-5-20250929`. The CUA runs inside a macOS virtual machine (Figure 3d), Lume [17], hosted on a local MacBook Air (Figure 3a). All click-through on-screen visualizations, including peripheral windows and annotations, are developed in Swift and deployed on both the macOS virtual machine (VM) and the MacBook host. For the real-time multimodal feedback described in Sections 4.2 and 4.3, including synchronized verbalizations, visualizations, and foley sounds, we use `gemini-2.5-flash` as a VLM-based converter that transforms plain-text CUA messages into SSML-like annotations [15]. The VLM is prompted to identify references to GUI objects and actions in Sidekick's messages and ground them in the current screenshot. For example, the CUA message in one of the proposed application scenarios (Figure 9):

> *"With the cat now selected with marching ants, the agent will click the generative fill button."*

can be converted into the following SSML-like annotation:

> *"With the* `<mark name="object_1"/>`[1] *cat* `</>`[2] *now selected with marching ants, the agent will* `<mark name="action_1"/>`[3] *click* `</>`[4] `<mark name="object_2"/>`[5] *the generative fill button* `</>`[6] *."*
>
> [1] `0.3 s`: Highlight the cat with a bounding box.
> [2] `1.1 s`: Remove the cat bounding box.
> [3] `3.9 s`: Display "click" and mark the click location.
> [4] `4.2 s`: Remove the action cue.
> [5] `4.2 s`: Highlight the button with a bounding box.
> [6] `5.6 s`: Remove the button bounding box.

Object labels are sent to a VLM (`gemini-3.5-flash`) for open-vocabulary object detection, which returns bounding boxes for visualization, while action labels specify when to render on-screen action cues and Foley sounds. The SSML-like annotation is then passed to Google TTS [10], which synthesizes speech and produces timestamps that Sidekick uses to synchronize verbal, visual, and auditory feedback.

5 Experiment

To understand the effectiveness of Sidekick in supporting multitasking with CUAs, we conducted a mixed-methods study with 30 participants, recruited via public posts at our institution (Mean age = 22.6, 18 female and 12 male), to quantitatively and qualitatively examine how Sidekick supports multitasking with CUAs (Table 1).

5.1 Tasks

To simulate multitasking, we carefully designed a primary and a secondary task based on prior work, pilot studies, and current CUA capabilities. Participants solved arithmetic problems as the **primary task** (e.g., two-digit addition or subtraction) while delegating a spreadsheet-filling task to the CUA as the **secondary task** (e.g., entering scores from a reference table; Figure 3d). We chose two-digit arithmetic tasks as they were commonly cited as cognitively demanding tasks in multitasking studies [27, 39, 42, 78]. Participants were placed in a scenario in which they served as a teaching assistant entering scores into an outdated grading system while managing an arithmetic assignment deadline, motivating them to seek help from CUAs. We selected spreadsheet-filling tasks due to the limited capabilities of current CUAs (mentioned in Section 3). Based on our empirical observations, CUAs were more controllable on these tasks and achieved higher success rates than on complex ones (e.g., slide or image editing).

For the secondary spreadsheet-filling task, the CUA was prompted to fill an entire column and then intentionally introduce two random errors to simulate unexpected behavior that could lead to serious mistakes (e.g., incorrectly failing students), requiring user attention. This was repeated across three columns (16 entries each). Each correct entry earned 3 points, incorrect entries incurred a 12-point penalty, and blanks received 0. Arithmetic answers were worth 1 point, with a 1-point penalty for errors and 0 for blanks. The spreadsheet task was weighted three times higher because it took human users about three times longer to complete. The 12-point penalty was carefully calibrated to encourage participants to monitor potential CUA errors: pilot studies showed that smaller penalties led participants to overlook errors, while larger penalties discouraged users from using CUA.

Note that Sidekick is not an agent or CUA, but a prototype system that serves as a communication layer for users to multitask with CUAs. It does not perform actions on the computer, but only provides feedback to users. Future work may explore integrating error remediation agents alongside CUAs to address errors automatically, which is beyond the scope of this paper.

5.2 Conditions and Research Questions

We included four multitasking conditions in this study:

- **(MN)** *Manual*: the user completes both tasks on their own.
- **(BL)** *Baseline*: a standard chat interface displaying only the CUA's textual messages.
- **(PT)** *Peripheral Text*: a standard chat interface, and a peripheral display presenting high-level summaries of the CUA's current states.
- **(SK)** *Sidekick*: a standard chat interface and a *Sidekick display*, as well as other multimodal feedback and summary.

We included **MN** to establish participants' baseline multitasking performance and to examine whether CUAs could improve performance compared to humans working alone. We included **BL** because chat-based interfaces currently represent the most common medium of communication with CUAs in existing products (e.g., Gemini [13], Operator [12]). **PT** incorporated textual feedback presented in a peripheral display, reflecting the design of prior ambient systems. Finally, **SK** included all Sidekick features described in the previous section. Across these four conditions, we investigated the following research questions:

- **RQ1** How do different feedback conditions affect users' multitasking performance when collaborating with a CUA?
- **RQ2** How do users allocate attention and interact with the CUA under different feedback conditions during multitasking?
- **RQ3** How do different feedback conditions influence users' perceived cognitive load, trust, and overall experience of multitasking with CUAs?

5.3 Interface Setup and Requirements

We provided participants with a MacBook Air running both Sidekick and a CUA in real time. Arithmetic tasks were presented in a Google Chrome interface, where users navigated via a sidebar or pressed *Enter* to submit answers and proceed (Figure 3a), with a visible session timer. A separate VM window displayed the CUA and

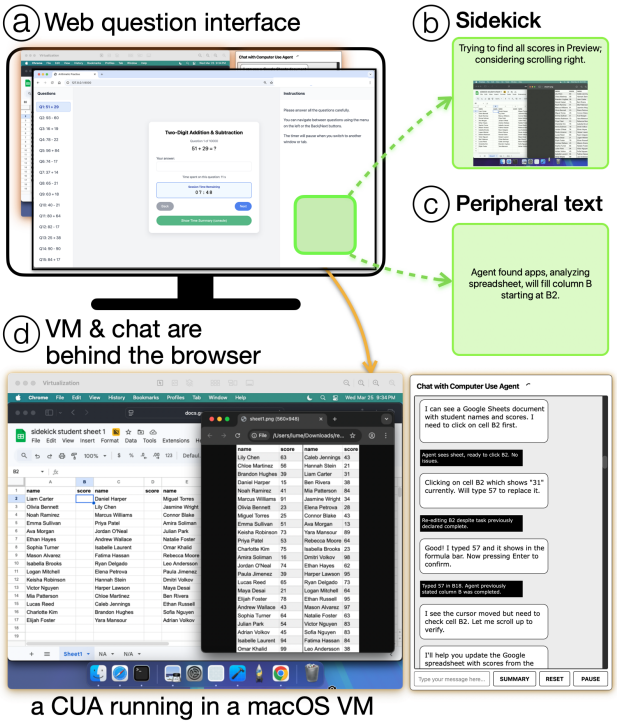

Figure 3: Study setup in a monitor. (a) Participants interacted with a browser for arithmetic questions, accompanied by a peripheral display for (b) *Sidekick* and (c) *Peripheral Text* conditions. (d) The CUA operated on a macOS virtual machine with its chat interface positioned behind the browser.

its chat interface, which included three controls: "RESET" (restart without context), "PAUSE" (temporarily halt while preserving context), and "SUMMARY" (exclusive to **SK**, providing on-demand summaries of the CUA's previous actions). The Chrome, VM and chat windows had fixed sizes, with the VM and chat positioned behind Chrome, and participants could not modify the layout. A movable peripheral display was available in **PT** and **SK** (Figure 3b,c). The VM window, peripheral display, and chat interface were linked: focusing on one brought the others to the foreground. An OS-level script ran in the background to record how long each window remained in focus for further analysis (Figure 5f).

5.4 Procedure

The study consisted of eight 8-minute sessions with sufficient rest provided between sessions (Figure 4). Participants were first introduced to all four multitasking conditions, including **MN**, **BL**, **PT**, and **SK**, and practice their delegated tasks. Then, participants began

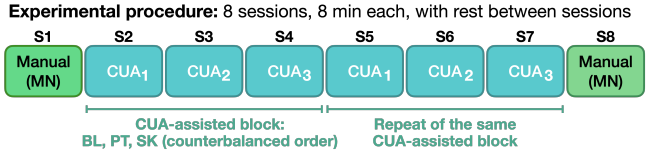

Figure 4: Experimental procedure. Participants completed eight 8-minute sessions: manual, two counterbalanced CUA-assisted blocks, and a final manual session.

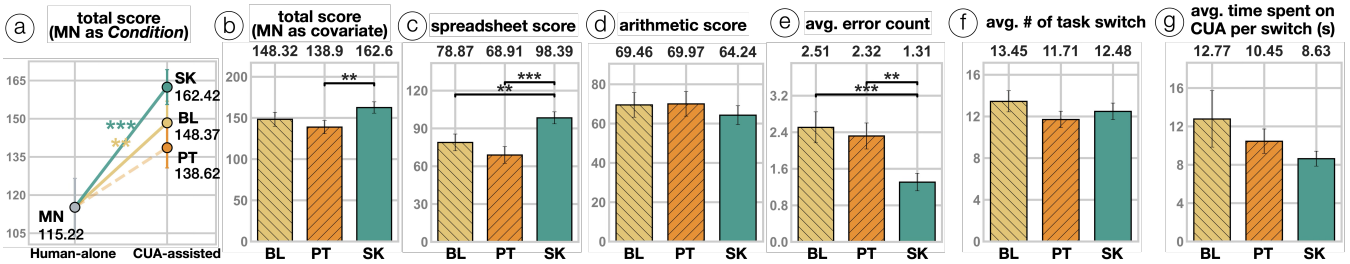

Figure 5: Results across conditions (MN, BL, PT, and SK) for task performance and interaction metrics. Panel (a) presents results from analyses treating MN as a fixed-effect condition alongside BL, PT, and SK. The remaining panels present results from analyses treating MN as a covariate: (b) *total score*, (c) *spreadsheet score*, (d) *arithmetic score*, (e) *average error count*, (f) *average number of task switches*, and (g) *average monitoring time per switch* for the CUA, in seconds. The mean scores in panels (a) and (b) slightly differ because the analyses modeled MN as a condition in panel (a) and as a covariate in panel (b). Bars represent means, and error bars indicate ±1 standard error. Statistical significance is denoted by $^{*}p < .05$, $^{**}p < .01$, and $^{***}p < .001$.

with the **MN** condition, followed by the other six sessions with three CUA-assisted conditions, including **BL**, **PT**, and **SK**, repeating twice and ended with another **MN** session. The order of three CUA-assisted conditions was counterbalanced across participants. The first and last **MN** sessions served to measure baseline performance and potential fatigue or learning effects. Participants were encouraged to maximize their scores, although their compensation was independent of task performance. After each session, participants completed Likert-scale questionnaires and the NASA-TLX [51] to assess cognitive load, after which they received feedback on their session scores. The study concluded with a post-study interview.

5.5 Dependent Measures and Data Analysis

For each session, we recorded task performance measures, including scores, error counts, task-switching behavior, and time spent across windows. We fitted two linear mixed-effects models. First, to compare CUA-assisted conditions (**BL**, **PT**, **SK**) with manual performance (**MN**), we included *Condition* as a fixed effect, *trial order* as a covariate, and participant ID as a random intercept. Second, to compare CUA-assisted conditions only, we fitted a model with *Condition* (**BL**, **PT**, **SK**) as a fixed effect and participant ID as a random intercept [65]. We also included *trial order* and *manual performance* (mean performance across the two **MN** sessions) as covariates to control for order effects and individual differences. Including *manual performance* as a covariate helped account for baseline ability, reducing the risk that observed differences were driven by individual skill rather than system effects and better isolating system-level gains. We did not observe fatigue or learning effects between the first and second **MN** conditions using paired t-tests and Wilcoxon signed-rank tests ($p > .16$). We did not include the first or second parts of the study as a covariate, as no significant effect was shown on performance ($\chi^2(1) = 1.15$, $p = .28$). Pairwise contrasts between conditions were tested using Wald tests with Holm correction for multiple comparisons.

5.6 Results

5.6.1 RQ1: How do different feedback conditions affect users' multitasking performance when collaborating with a CUA?. **Human-CUA collaboration improves multitasking performance over human alone, and Sidekick's multimodal feedback further enhances this collaboration without disrupting users' primary task.** We evaluated multitasking performance using three metrics: *total score*, *arithmetic score*, and *spreadsheet score*, where *total score* is the sum of the other two (Figure 5). We first compared **MN** with CUA-assisted conditions (including BL, PT, and SK) to assess whether human-CUA collaboration improves *total score* over a human alone (Figure 5a). A linear mixed-effects model revealed a significant main effect of *Condition* on *total score* (Wald $\chi^2(3) = 23.03$, $p < .001$). The estimated marginal means were 115.22 for **MN**, 148.37 for **BL**, 138.62 for **PT**, and 162.42 for **SK**. Holm-corrected pairwise comparisons showed that both **BL** ($p = .007$) and **SK** ($p < .001$) significantly outperformed **MN**. Although **PT** also achieved a higher score than **MN**, the difference was not significant after correction ($p = .075$), suggesting that CUAs are not necessarily beneficial when their feedback is not well designed.

We next compared *total score* across the CUA-assisted conditions (**BL**, **PT**, and **SK**; Figure 5b). To account for individual differences in general task ability, we included each participant's performance in **MN** as a covariate in the linear mixed-effects model. A significant main effect of *Condition* was observed (Wald $\chi^2(2) = 12.31$, $p = .002$), with **SK** (M=162.6) outperforming **PT** (M=138.9, $p = .001$) and showing a marginal trend over **BL** (M=148.32, $p = .079$).

To understand the source of this improvement, we examined the two task components separately (Figure 5c, d). We found that the effect of *Condition* on *total score* was driven by *spreadsheet score* (Wald $\chi^2(2) = 21.52$, $p < .001$), where **SK** (M=98.39) significantly outperformed both **PT** (M=68.90, $p < .001$) and **BL** (M=78.87, $p = .006$). In contrast, *arithmetic score* showed no significant differences (Wald $\chi^2(2) = 3.01$, $p = .222$). This indicated that the different feedback mechanisms used to communicate with the CUA did not interfere with users' primary arithmetic task performance.

These results suggest that *human-CUA collaboration enables users to achieve higher performance than human alone, but the communication mechanisms must be carefully designed.* For instance, **PT** did not outperform **BL**, as peripheral text and static color provided weak signals of the CUA's state. Furthermore, the unchanging color might create a false sense of security, leading users to disengage and feel it *"made it seem everything was fine"* (P9). Also, text feedback required additional effort to interpret, *"without color changes, it takes extra time to read the text and figure out if something is wrong"*

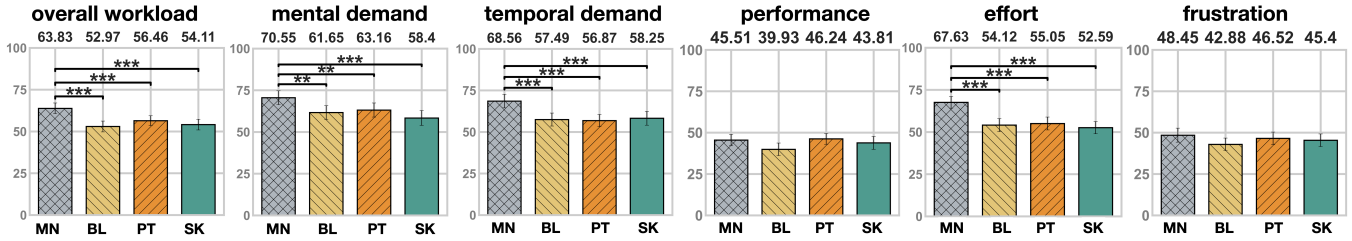

Figure 6: NASA-TLX responses across conditions (MN, BL, PT, and SK) for *overall workload*, *mental demand*, *temporal demand*, *performance* (the higher, the better), *effort*, and *frustration*. We excluded *physical demand* because it was not relevant to the study context. Bars show mean values with error bars indicating ±1 standard error. Statistical significance is indicated as $p < .05$ (*), $p < .01$ (**), and $p < .001$ (***).

(P15), and was sometimes overlooked as *"it was impossible to do the calculation and see the text at the same time"* (P17).

In contrast, *Sidekick's multimodal feedback improved collaboration without disrupting the primary task.* However, preferences on multimodal feedback varied: some participants favored a simpler design (**BL**), noting that the rich information in **SK** could be distracting. As P10 noted, *"Too many things going on split my attention and threw me off. I preferred the baseline for simplicity. I'd rather have a simple, on-demand color cue."* This attentional fragmentation may explain why **SK** showed only marginal improvement over **BL**. We next analyze user behaviors across conditions to better understand how these performance differences emerged.

5.6.2 RQ2: How do users allocate attention and interact with the CUA under different feedback conditions during multitasking? **Sidekick enabled more effective intervention when errors occurred, but did not require more attention from users.**

To better understand the performance differences observed in **RQ1**, we analyzed *spreadsheet error counts* and user interaction behaviors, including *monitoring time* and *task switching*.

A linear mixed-effects model revealed a significant effect of *Condition* on *spreadsheet error counts* (Wald $\chi^2(2) = 15.46$, $p < .001$). Estimated marginal means were 2.51 errors for **BL**, 2.32 for **PT**, and 1.31 for **SK** (Figure 5e). Holm-corrected post-hoc comparisons showed that **SK** produced significantly fewer errors than both **BL** ($p < .001$) and **PT** ($p = .004$), while the difference between **BL** and **PT** was not significant. This suggested that Sidekick helped users address CUA errors better than **BL** and **PT** in the spreadsheet task.

We then examine whether these improvements resulted from increased monitoring of the CUA by analyzing the *number of task switches* (Figure 5g), and the *time spent* monitoring the spreadsheet task (Figure 5f). No significant effects were found for *task switching* (Wald $\chi^2(2) = 5.06$, $p = .080$) or *monitoring time* (Wald $\chi^2(2) = 2.56$, $p = .278$). Users switched tasks a similar number of times (**BL**: 13.34, **PT**: 11.70, **SK**: 12.48) and spent comparable time monitoring (**BL**: 12.78s, **PT**: 10.45s, **SK**: 8.63s).

These results suggest that *Sidekick improved collaboration without increasing monitoring effort.* Qualitative feedback further supported this finding. Participants (N=21) noted that color cues in the *Sidekick display* clearly signaled the CUA's state and enabled timely intervention. As P17 noted, *"**SK** was easy ... I didn't have to focus on both tasks ... I could tell from colors like yellow or red and step in before things got worse."* It also reduced the need to review chat logs; as P10 explained, *"it takes time to go through logs in **BL**, but color cues help me quickly understand what's happening and make decisions."* Other visual cues further improved interpretability. Thumbnails (N=12) and high-level summaries of CUA states (N=19) in *Sidekick display* provided contextual grounding, making it *"easier to tell what the agent is doing than text alone [**PT**]"* (P6) or *"immediately see how far it has gotten"* (P2), while recency indicators (N=13) helped track recent actions without scanning histories, *"seeing where it was working most recently, so I don't have to scroll through the chat"* (P24).

Together, these feedback explain the improved CUA-assisted task performance in **RQ1**, enabling users to address errors timely without increased monitoring or extensive browsing of chat logs compared to **BL** and **PT**.

5.6.3 RQ3: How do different feedback conditions influence users' perceived cognitive load, trust, and overall experience of multitasking with CUAs? **Human-CUA collaboration significantly reduced perceived workload compared to human alone. Sidekick's multimodal feedback was perceived as more helpful for supporting multitasking and enabling timely intervention.**

While *RQ1* and *RQ2* examined objective performance and interaction behaviors, we next assessed participants' subjective experience using NASA-TLX and post-task Likert ratings.

NASA-TLX results showed that human–CUA collaboration significantly reduced perceived workload compared to working alone (Figure 6). A linear mixed-effects model revealed a main effect of *Condition* (Wald $\chi^2(3)$, $p < .001$), with higher workload in **MN** (M=63.83) than in **BL** (M=52.97), **PT** (M=56.46), and **SK** (M=54.11). Post-hoc comparisons confirmed that all CUA-assisted conditions were significantly lower than **MN**, with no differences among **BL**, **PT**, and **SK**. Similar patterns were observed for *mental demand*, *temporal demand*, and *effort*, while perceived *performance* and *frustration* showed no differences.

Qualitative feedback further supported these findings, where the decreased workload under CUA-assisted conditions was primarily due to task delegation, with users shifting from the role of execution to monitoring. For example, P1 noted that the CUA could achieve *"roughly 80% accuracy,"* transferring their role to *"verifying revised cells rather than checking every cell entry."* Similarly, P18 highlighted that the spreadsheet-filling tasks were *"time-consuming, and offloading them to the agent significantly reduced effort."* However, monitoring introduced additional effort, as P26 explained, *"you still*

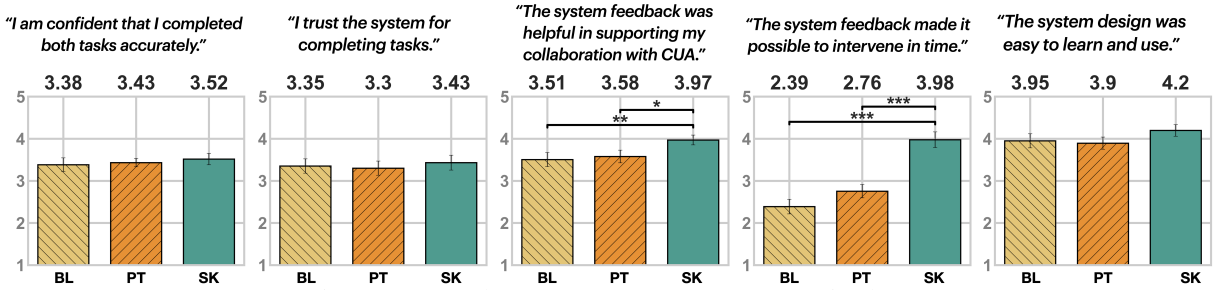

Figure 7: Results across conditions (BL, PT, and SK) for five Likert-scale measures (1–5): confidence in completing both tasks accurately, trust in the system for task completion, perceived helpfulness of system feedback for multitasking with the CUA, perceived ability to notice errors in time to intervene, and perceived ease of learning and use. Bars show mean values with error bars indicating ±1 standard error. Statistical significance is indicated as $p < .05$ (*), $p < .01$ (**), and $p < .001$ (***).

have to check the agent... monitoring becomes an extra task", which may explain the similar workload across CUA-assisted conditions.

Post-task Likert ratings showed similar trust, confidence, and usability across **BL**, **PT**, and **SK** (Figure 7). Perceived trust and confidence were comparable across conditions, likely influenced by the imperfect CUA, which was intentionally designed to introduce errors that required participants to monitor and intervene. Similarly, no significant differences were found in ease of understanding or use, which is expected given that the feedback mechanisms in **BL** (chat-only) and **PT** (peripheral display) were also relatively simple.

However, participants rated **SK** as significantly more helpful for multitasking than both **BL** ($p = .005$) and **PT** ($p = .020$), and as enabling more timely intervention when errors occurred (both $p < .001$). Qualitative feedback supports these findings, especially for transition support and transparency. The *Sidekick display* reduced context switching, as P11 noted, *"it reduced the effort of switching to the background to see how it's running compared to the **BL**."* Upon resuming tasks with the CUA, Sidekick helped users quickly understand prior actions; as P22 noted, *"replay was good for seeing what the agent edited. I usually check the cell it just finished."* Verbal feedback also maintained awareness during transitions, with P23 explaining, *"when I switched back to the math window, I could still hear what it did, making sure it worked properly after I left."* Finally, real-time visualization of the CUA's reasoning improved transparency; as P28 noted, *"the animation was helpful because I could see what's happening. I feel more confident that the answers are real if I can see how they were derived."*

Together, these results suggested that human–CUA collaboration reduced workload by offloading execution, but still required monitoring, leading to similar workload across CUA conditions. Importantly, Sidekick's multimodal feedback did not increase cognitive burden and was perceived as more effective for timely intervention, transparency, and rapid context resumption.

6 Proposed Application Scenarios

We propose three applications to demonstrate how Sidekick's design principles and interactions could extend beyond spreadsheet tasks to broader workflows. We acknowledge, however, that their benefits have not yet been systematically evaluated, and we view such evaluation as an important direction for future work.

Supporting complex multimodal tasks (Figure 8). Future CUAs may enable extended workflows spanning multiple applications. For example, a user might ask a CUA to prepare a presentation by gathering content, inserting charts, and formatting layouts, producing outputs that require timely aesthetic judgment. In such settings, thumbnails in the *Sidekick display* could enhance awareness of images, layouts, and styles, supporting timely intervention and informed decisions. Color cues could extend beyond error signaling to reflect the impact of design changes (Figure 8a; *e.g.*, low impact: font changes; high impact: overall style). Upon return, Sidekick could summarize prior actions through verbal explanations and replay, highlighting design rationale, edited elements, and layout changes (Figure 8b). These capabilities could generalize to other creative multimodal workflows, such as image or video editing.

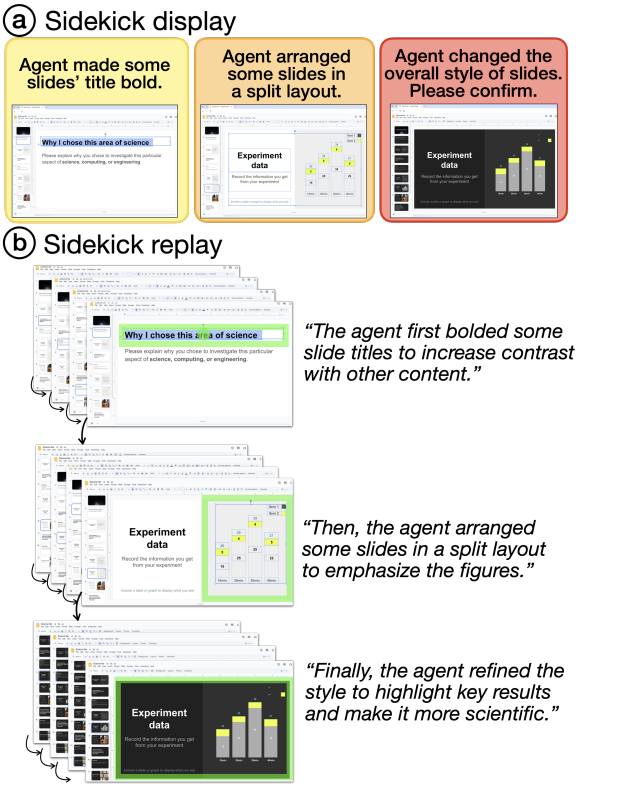

Figure 8: Sidekick's potential application in slide editing. (a) Color cues in the *Sidekick display* can encode the impact of design changes. (b) Sidekick can replay prior visual changes and explain design rationale upon resumption.

Supporting new interface learning over the CUA's and Sidekick's shoulder (Figure 9). Learning new software interfaces has long been studied in HCI, with prior work leveraging tutorials from videos and crowd-sourced sources (e.g., blogs) [56, 62–64, 68, 70, 90]. As AI increasingly enables users to generate applications and interfaces in natural language, learning new software or interfaces developed by others is likely to become more common. Sidekick offers a learning opportunity when integrated with CUAs: by providing real-time visualizations and verbal feedback of CUA reasoning, it enables users to observe and learn from the agent's actions. For example, when prompted with a tutorial on features like "Generative Fill," Sidekick can explain steps sourced from the online community while the CUA demonstrates them in real time (Figure 9).

Figure 9: Sidekick's potential application in software tutorials. Sidekick could explain how to achieve a function step by step while the CUA demonstrates the actions. Users could observe and learn from the real-time visualizations and verbal feedback.

Reporting usability tests of generative UIs to users (Figure 10). GUI testing is essential for usability and accessibility [73, 83, 85]. Agentic platforms can support closed-loop workflows where coding agents generate GUIs and CUAs evaluate them against heuristic or accessibility guidelines [4]. However, communication between agents and users remains limited. Sidekick could bridge this gap by surfacing usability issues in real time through its display or providing detailed multimodal reports upon user engagement (Figure 10a). It may also track recurring errors or usability issues and summarize them for users after several iterations. Users can intervene by monitoring the *Sidekick display*, where color cues reflect issue severity (Figure 10b), signaling when prompt revisions are needed. Additionally, summaries upon return, along with real-time visualization and verbal feedback, can help users verify testing rationale against intended guidelines (Figure 10c), improving transparency and confidence.

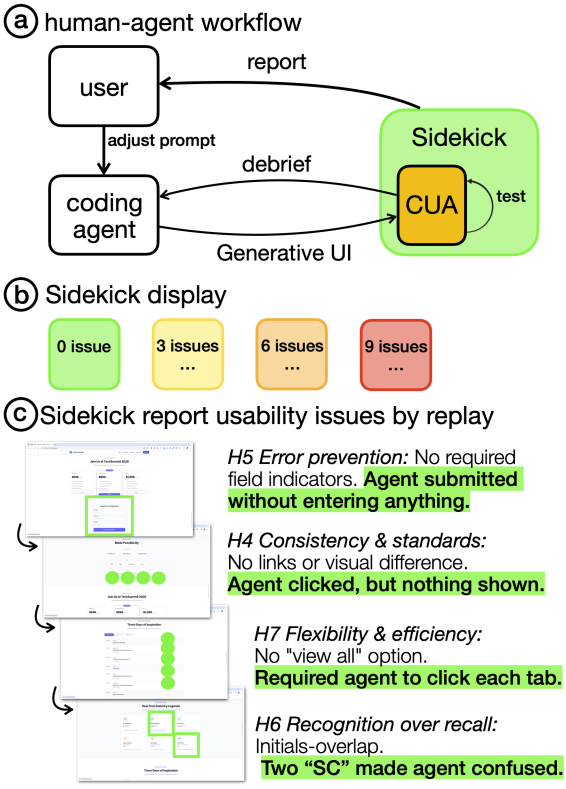

Figure 10: Sidekick's potential application in collaboration with coding agents and CUAs for a usability test report. (a) Integration into closed-loop workflows for UI generation and evaluation. (b) Color cues in the *Sidekick display* encode the number or severity of identified usability issues. (c) Sidekick reports key results by explaining issues alongside relevant guidelines.

7 Discussion and Future Work

We discuss the limitations, our lessons learned, and implications for extending Sidekick in broader scenarios and future CUAs.

Design implications for multimodal human-CUA communication. In our study, Sidekick improved multitasking with CUAs over baseline systems via multimodal feedback to communicate agent states. Recent work by Cheng et al. [37] proposed a taxonomy of human-CUA interactions; Sidekick partially aligns with this taxonomy by making agent activities visible, increasing transparency of reasoning, communicating runtime status, and enabling user intervention when errors occur. However, feedback design still requires additional user-centered refinement. For example, preferences for feedback modality varied: some participants found visual feedback (e.g., color changes) helpful but distracting (Section 5.6.1), and suggested lighter alternatives such as brief audio cues. In contrast, in foreground interactions, some noted that visual feedback alone may suffice, while speech was unnecessary. These findings highlighted the need for customizable feedback and context awareness, which could draw on prior design principles or systems in attention management, such as tuning information density to manageable chunks [32, 34, 43, 80, 89], adapting notification modality

to user context [22, 26, 30, 31, 67, 82], and delivering information at opportune moments [23, 36, 55, 57, 58]. Finally, beyond conveying CUAs' operation states, future work could further facilitate bidirectional human-CUA alignment by communicating user contexts to CUAs (e.g., goals, safety, and privacy constraints) and clarifying CUAs' capabilities and knowledge scope to users [21, 37].

Supporting human-agent communication at scale. In this work, we developed Sidekick to address the communication gap between humans and CUAs during multitasking, with a focus on enhancing error awareness in our study. Rather than surfacing every error, Sidekick acts as a monitoring proxy that requests user intervention only when accumulated errors exceed a threshold (Section 4.1). This enables strategic check-ins while allowing CUAs to recover independently from minor issues.

However, as CUAs become faster and more accurate, users' communication needs may shift beyond error detection toward subjective concerns such as aesthetics, preferences, and privacy. In such cases, Sidekick could therefore support richer interactions that communicate nuanced decisions and outcomes, as our application scenarios demonstrated (Figure 8). As CUA speed increases, users may also face a surge of outputs in a short time. To address this, Sidekick could provide hierarchical summaries that first surface high-stakes artifacts (e.g., key slides among many outputs) or information likely to contain errors or expectation mismatches, and allow users to drill down into other details if needed.

Moreover, Sidekick currently focuses on a single CUA performing one task within one window, whereas real-world settings may involve multiple agents operating across different workspaces as such systems become more capable and widespread. This direction aligns with prior work on unified interfaces for managing multiple workspaces [25, 29, 59, 60]. Accordingly, multi-workspace environments may become the primary context for Sidekick, supporting coordination across multiple agents. The Sidekick display could shift from per-agent views to aggregated summaries (e.g., overall progress indicators or selective visual snippets highlighting issues [86]), with color cues encoding cross-agent states such as conflicts or dependencies. Upon returning, Sidekick could provide selective summaries of cross-agent activities, prioritizing information based on user goals. In foreground scenarios, it could highlight critical agents or visualize dependencies to help users detect conflicts and coordinate interventions.

Beyond supporting users during task execution, recent agent trajectory-analysis systems have transformed lengthy execution traces into behavioral categories, diagnostic measures, and visual representations for post-hoc inspection [11, 48, 61, 74]. Building on these approaches, Sidekick could be extended to translate low-level human-CUA activities into structured reports that surface consequential behaviors, recurring risks and errors, the effects of user interventions, and moments when additional intervention may be warranted. Such reports could help identify systematic failure patterns and inform improvements to monitoring and intervention strategies. More broadly, in the future, Sidekick could evolve into a scalable communication layer for increasingly complex and capable multi-agent workflows, providing real-time feedback tailored to users' goals, attention, and workload, alongside post-hoc reports that support retrospective analysis, auditing, and iterative system improvement.

Study limitations. As noted in Section 3, CUAs at the time of our study had limited reliability, leading us to use more controlled tasks (e.g., spreadsheet filling). These tasks reflect common use cases observed in commercial demonstrations [1] and our formative study, which *users can perform the work themselves but prefer to delegate it due to its tedious and repetitive nature.* While CUA capabilities are expected to improve, our findings may still generalize to similar repetitive tasks (e.g., tax filing or slide formatting), as noted by participants. Additionally, although reward and penalty mechanisms were calibrated through pilot studies, they may have influenced participants' mental models and strategies. However, we did not observe strong evidence of altered interaction patterns that impact study results (e.g., abandoning CUA outputs or ignoring errors). Future work could examine tasks with varying risk levels or alternative incentive structures to better understand their effects on user behavior, trust, and confidence.

8 Conclusion

In this work, we introduce Sidekick, a prototype system that supports human-CUA communication during multitasking. Sidekick provides peripheral signals for background awareness, structured multimodal summaries for context resumption, and real-time synchronized feedback for foreground transparency. Through a mixed-methods study with 30 participants, we show that Sidekick significantly improves multitasking performance over working alone and text-based baselines by enhancing awareness of CUA progress and enabling timely intervention without constant monitoring or disrupting primary tasks. We further demonstrate its potential in broader scenarios, including multimodal task support, reporting usability issues in multi-agent workflows, and enabling learning through agent demonstrations. Finally, we discuss implications for more customizable, lightweight human–CUA communication and the scalability of Sidekick as CUAs continue to improve.

Acknowledgments

We thank all study participants and anonymous reviewers for their valuable feedback, as well as Prof. Xu Wang for suggestions on the experimental design and analysis. This research was supported in part by an Adobe Research Gift and a Google Academic Research Award. Ruei-Che Chang was also supported by the Apple Scholars in AI/ML PhD Fellowship.

References

[1] 2024. Introducing computer use, a new Claude 3.5 Sonnet, and Claude 3.5 Haiku. https://www.anthropic.com/news/3-5-models-and-computer-use
[2] 2026. Anthropic. https://www.anthropic.com/
[3] 2026. Claude Code. https://claude.com/product/claude-code
[4] 2026. Claude Cowork by Anthropic. https://www.anthropic.com/product/claude-cowork
[5] 2026. Computer use. https://platform.openai.com/docs/guides/tools-computer-use
[6] 2026. Cua. https://cua.ai/
[7] 2026. Cursor. https://cursor.com/
[8] 2026. Developing a computer use model. https://www.anthropic.com/news/developing-computer-use
[9] 2026. Foley (sound design). https://en.wikipedia.org/wiki/Foley_(sound_design)
[10] 2026. Google Text-to-Speech AI. https://cloud.google.com/text-to-speech
[11] 2026. Introducing Docent. https://transluce.org/introducing-docent
[12] 2026. Introducing Operator. https://openai.com/index/introducing-operator/
[13] 2026. Introducing the Gemini 2.5 Computer Use model. https://blog.google/technology/google-deepmind/gemini-computer-use-model/

[14] 2026. OSWorld: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments. https://os-world.github.io/

[15] 2026. Speech Synthesis Markup Language (SSML). https://docs.cloud.google.com/text-to-speech/docs/ssml

[16] 2026. Vercept. https://vercept.com/

[17] 2026. What is Lume? https://cua.ai/docs/lume/guide/getting-started/introduction

[18] Reyna Abhyankar, Qi Qi, and Yiying Zhang. 2025. Osworld-human: Benchmarking the efficiency of computer-use agents. *arXiv preprint arXiv:2506.16042* (2025).

[19] Piotr D. Adamczyk and Brian P. Bailey. 2004. If not now, when? the effects of interruption at different moments within task execution. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vienna, Austria) *(CHI '04)*. Association for Computing Machinery, New York, NY, USA, 271–278. https://doi.org/10.1145/985692.985727

[20] Saleema Amershi, Max Chickering, Steven M. Drucker, Bongshin Lee, Patrice Simard, and Jina Suh. 2015. ModelTracker: Redesigning Performance Analysis Tools for Machine Learning. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 337–346. https://doi.org/10.1145/2702123.2702509

[21] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300233

[22] Ernesto Arroyo, Ted Selker, and Alexandre Stouffs. 2002. Interruptions as multimodal outputs: Which are the less disruptive?. In *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*. IEEE, 479–482.

[23] James "Bo" Begole, Nicholas E. Matsakis, and John C. Tang. 2004. Lilsys: Sensing Unavailability. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work* (Chicago, Illinois, USA) *(CSCW '04)*. Association for Computing Machinery, New York, NY, USA, 511–514. https://doi.org/10.1145/1031607.1031691

[24] Rogerio Bonatti, Dan Zhao, Francesco Bonacci, Dillon Dupont, Sara Abdali, Yinheng Li, Yadong Lu, Justin Wagle, Kazuhito Koishida, Arthur Bucker, et al. 2024. Windows agent arena: Evaluating multi-modal os agents at scale. *arXiv preprint arXiv:2409.08264* (2024).

[25] Andrew Bragdon, Robert Zeleznik, Steven P. Reiss, Suman Karumuri, William Cheung, Joshua Kaplan, Christopher Coleman, Ferdi Adeputra, and Jr. LaViola, Joseph J. 2010. Code bubbles: a working set-based interface for code understanding and maintenance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 2503–2512. https://doi.org/10.1145/1753326.1753706

[26] Stephen A. Brewster, Peter C. Wright, and Alistair D. N. Edwards. 1994. The design and evaluation of an auditory-enhanced scrollbar. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, Massachusetts, USA) *(CHI '94)*. Association for Computing Machinery, New York, NY, USA, 173–179. https://doi.org/10.1145/191666.191733

[27] Valdimar Briem and Leif R Hedman. 1995. Behavioural effects of mobile telephone use during simulated driving. *Ergonomics* 38, 12 (1995), 2536–2562.

[28] J. J. Cadiz, Gina Venolia, Gavin Jancke, and Anoop Gupta. 2002. Designing and deploying an information awareness interface. In *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work* (New Orleans, Louisiana, USA) *(CSCW '02)*. Association for Computing Machinery, New York, NY, USA, 314–323. https://doi.org/10.1145/587078.587122

[29] Joseph Chee Chang, Yongsung Kim, Victor Miller, Michael Xieyang Liu, Brad A Myers, and Aniket Kittur. 2021. Tabs.do: Task-Centric Browser Tab Management. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21)*. Association for Computing Machinery, New York, NY, USA, 663–676. https://doi.org/10.1145/3472749.3474777

[30] Ruei-Che Chang, Tovi Grossman, Carine Rognon, Michael Glueck, Christopher Collins, Amy Karlson, and Hemant Bhaskar Surale. 2025. Viago: Exploring Visual-Audio Modality Transitions for Social Media Consumption on the Go. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25)*. Association for Computing Machinery, New York, NY, USA, Article 20, 20 pages. https://doi.org/10.1145/3746059.3747761

[31] Ruei-Che Chang, Chia-Sheng Hung, Bing-Yu Chen, Dhruv Jain, and Anhong Guo. 2024. SoundShift: Exploring Sound Manipulations for Accessible Mixed-Reality Awareness. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (Copenhagen, Denmark) *(DIS '24)*. Association for Computing Machinery, New York, NY, USA, 116–132. https://doi.org/10.1145/3643834.3661556

[32] Ruei-Che Chang, Yuxuan Liu, and Anhong Guo. 2024. WorldScribe: Towards Context-Aware Live Visual Descriptions. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 140, 18 pages. https://doi.org/10.1145/3654777.3676375

[33] Ruei-Che Chang, Yuxuan Liu, Lotus Zhang, and Anhong Guo. 2024. EditScribe: Non-Visual Image Editing with Natural Language Verification Loops. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility* (St. John's, NL, Canada) *(ASSETS '24)*. Association for Computing Machinery, New York, NY, USA, Article 65, 19 pages. https://doi.org/10.1145/3663548.3675599

[34] Ruei-Che Chang, Rosiana Natalie, Wenqian Xu, Jovan Zheng Feng Yap, Tiange Luo, Venkatesh Potluri, and Anhong Guo. 2026. TouchScribe: Augmenting Non-Visual Hand-Object Interactions with Automated Live Visual Descriptions. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. Association for Computing Machinery, New York, NY, USA, Article 747, 18 pages. https://doi.org/10.1145/3772318.3791308

[35] Jessie YC Chen, Shan G Lakhmani, Kimberly Stowers, Anthony R Selkowitz, Julia L Wright, and Michael Barnes. 2018. Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical issues in ergonomics science* 19, 3 (2018), 259–282.

[36] Kuan-Wen Chen, Yung-Ju Chang, and Liwei Chan. 2022. Predicting Opportune Moments to Deliver Notifications in Virtual Reality. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 186, 18 pages. https://doi.org/10.1145/3491102.3517529

[37] Ruijia Cheng, Jenny T Liang, Eldon Schoop, and Jeffrey Nichols. 2026. Mapping the Design Space of User Experience for Computer Use Agents. In *Proceedings of the 31st International Conference on Intelligent User Interfaces (IUI '26)*. ACM, 646–662. https://doi.org/10.1145/3742413.3789132

[38] Yi Fei Cheng, Jarod Bloch, Alexander Wang, Andrea Bianchi, Anusha Withana, Anhong Guo, Laurie M. Heller, and David Lindlbauer. 2026. Auditorily Embodied Conversational Agents: Effects of Spatialization and Situated Audio Cues on Presence and Social Perception. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. Association for Computing Machinery, New York, NY, USA, Article 1465, 16 pages. https://doi.org/10.1145/3772318.3791794

[39] Yi Fei Cheng, Hirokazu Shirado, and Shunichi Kasahara. 2025. Conversational Agents on Your Behalf: Opportunities and Challenges of Shared Autonomy in Voice Communication for Multitasking. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 160, 18 pages. https://doi.org/10.1145/3706598.3714017

[40] Dennis Collaris and Jarke J Van Wijk. 2020. ExplainExplore: Visual exploration of machine learning explanations. In *2020 IEEE Pacific Visualization Symposium (PacificVis)*. IEEE, 26–35.

[41] Sunny Consolvo and Jeffrey Towle. 2005. Evaluating an ambient display for the home. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems* (Portland, OR, USA) *(CHI EA '05)*. Association for Computing Machinery, New York, NY, USA, 1304–1307. https://doi.org/10.1145/1056808.1056902

[42] Mary Czerwinski, Edward Cutrell, and Eric Horvitz. 2000. Instant messaging and interruption: Influence of task type on performance. In *OZCHI 2000 conference proceedings*, Vol. 356. 361–367.

[43] Laura Dabbish and Robert E. Kraut. 2004. Controlling interruptions: awareness displays and social motivation for coordination. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work* (Chicago, Illinois, USA) *(CSCW '04)*. Association for Computing Machinery, New York, NY, USA, 182–191. https://doi.org/10.1145/1031607.1031638

[44] Anindya Das Antar, Somayeh Molaei, Yan-Ying Chen, Matthew L Lee, and Nikola Banovic. 2024. VIME: Visual Interactive Model Explorer for Identifying Capabilities and Limitations of Machine Learning Models for Sequential Decision-Making. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 59, 21 pages. https://doi.org/10.1145/3654777.3676323

[45] Edward S. De Guzman, Margaret Yau, Anthony Gagliano, Austin Park, and Anind K. Dey. 2004. Exploring the design and use of peripheral displays of awareness information. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems* (Vienna, Austria) *(CHI EA '04)*. Association for Computing Machinery, New York, NY, USA, 1247–1250. https://doi.org/10.1145/985921.986035

[46] Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. 2023. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems* 36 (2023), 28091–28114.

[47] Anton N. Dragunov, Thomas G. Dietterich, Kevin Johnsrude, Matthew McLaughlin, Lida Li, and Jonathan L. Herlocker. 2005. TaskTracer: a desktop environment to support multi-tasking knowledge workers. In *Proceedings of the 10th International Conference on Intelligent User Interfaces* (San Diego, California, USA) *(IUI '05)*. Association for Computing Machinery, New York, NY, USA, 75–82. https://doi.org/10.1145/1040830.1040855

[48] Jie Gao, Kaiser Sun, Jen tse Huang, Katherine Van Koevering, Sijie Ji, Heyuan Huang, Weiyan Shi, Zhuoran Lu, Ziang Xiao, Daniel Khashabi, and Mark Dredze. 2026. How to Interpret Agent Behavior. arXiv:2605.13625 [cs.AI] https://arxiv.

org/abs/2605.13625

[49] Boyu Gou, Zanming Huang, Yuting Ning, Yu Gu, Michael Lin, Weijian Qi, Andrei Kopanev, Botao Yu, Bernal Jiménez Gutiérrez, Yiheng Shu, et al. 2025. Mind2web 2: Evaluating agentic search with agent-as-a-judge. *arXiv preprint arXiv:2506.21506* (2025).

[50] Tovi Grossman, Justin Matejka, and George Fitzmaurice. 2010. Chronicle: capture, exploration, and playback of document workflow histories. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology* (New York, New York, USA) *(UIST '10)*. Association for Computing Machinery, New York, NY, USA, 143–152. https://doi.org/10.1145/1866029.1866054

[51] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage publications Sage CA: Los Angeles, CA, 904–908.

[52] Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan, and Dong Yu. 2024. Webvoyager: Building an end-to-end web agent with large multimodal models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 6864–6890.

[53] Fred Hohman, Arjun Srinivasan, and Steven M Drucker. 2019. TeleGam: Combining visualization and verbalization for interpretable machine learning. In *2019 ieee visualization conference (VIS)*. IEEE, 151–155.

[54] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, Pennsylvania, USA) *(CHI '99)*. Association for Computing Machinery, New York, NY, USA, 159–166. https://doi.org/10.1145/302979.303030

[55] Eric Horvitz, Paul Koch, and Johnson Apacible. 2004. BusyBody: creating and fielding personalized models of the cost of interruption. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work* (Chicago, Illinois, USA) *(CSCW '04)*. Association for Computing Machinery, New York, NY, USA, 507–510. https://doi.org/10.1145/1031607.1031690

[56] Nathaniel Hudson, Benjamin Lafreniere, Parmit K. Chilana, and Tovi Grossman. 2018. Investigating How Online Help and Learning Resources Support Children's Use of 3D Design Software. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173831

[57] Shamsi T. Iqbal and Brian P. Bailey. 2008. Effects of intelligent notification management on users and their tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 93–102. https://doi.org/10.1145/1357054.1357070

[58] Shamsi T. Iqbal and Brian P. Bailey. 2011. Oasis: A framework for linking notification delivery to the perceptual structure of goal-directed tasks. *ACM Trans. Comput.-Hum. Interact.* 17, 4, Article 15 (Dec. 2011), 28 pages. https://doi.org/10.1145/1879831.1879833

[59] Peiling Jiang and Haijun Xia. 2025. Orca: Browsing at scale through user-driven and ai-facilitated orchestration across malleable webpages. *arXiv preprint arXiv:2505.22831* (2025).

[60] Eser Kandogan and Ben Shneiderman. 1997. Elastic Windows: evaluation of multi-window operations. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*. 250–257.

[61] Wonjoong Kim, Sangwu Park, Yeonjun In, Sein Kim, Dongha Lee, and Chanyoung Park. 2026. Beyond the Final Answer: Evaluating the Reasoning Trajectories of Tool-Augmented Agents. arXiv:2510.02837 [cs.AI] https://arxiv.org/abs/2510.02837

[62] Benjamin Lafreniere, Tovi Grossman, and George Fitzmaurice. 2013. Community enhanced tutorials: improving tutorials with multiple demonstrations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1779–1788.

[63] Ben Lafreniere, Tovi Grossman, Justin Matejka, and George Fitzmaurice. 2014. Investigating the feasibility of extracting tool demonstrations from in-situ video content. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 4007–4016. https://doi.org/10.1145/2556288.2557142

[64] Wei Li, Tovi Grossman, and George Fitzmaurice. 2012. GamiCAD: a gamified tutorial system for first time autocad users. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology* (Cambridge, Massachusetts, USA) *(UIST '12)*. Association for Computing Machinery, New York, NY, USA, 103–112. https://doi.org/10.1145/2380116.2380131

[65] Mary J Lindstrom and Douglas M Bates. 1988. Newton—Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *J. Amer. Statist. Assoc.* 83, 404 (1988), 1014–1022.

[66] Blair MacIntyre, Elizabeth D. Mynatt, Stephen Voida, Klaus M. Hansen, Joe Tullio, and Gregory M. Corso. 2001. Support for multitasking and background awareness using interactive peripheral displays. In *Proceedings of the 14th Annual ACM Symposium on User Interface Software and Technology* (Orlando, Florida) *(UIST '01)*. Association for Computing Machinery, New York, NY, USA, 41–50. https://doi.org/10.1145/502348.502355

[67] Paul P. Maglio and Christopher S. Campbell. 2000. Tradeoffs in displaying peripheral information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (The Hague, The Netherlands) *(CHI '00)*. Association for Computing Machinery, New York, NY, USA, 241–248. https://doi.org/10.1145/332040.332438

[68] Justin Matejka, Tovi Grossman, and George Fitzmaurice. 2011. Ambient help. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 2751–2760. https://doi.org/10.1145/1978942.1979349

[69] Justin Matejka, Tovi Grossman, and George Fitzmaurice. 2021. MeetingMate: an Ambient Interface for Improved Meeting Effectiveness and Corporate Knowledge Sharing. In *Graphics Interface 2021*. https://openreview.net/forum?id=ibaCpFUWVb9

[70] Justin Matejka, Wei Li, Tovi Grossman, and George Fitzmaurice. 2009. CommunityCommands: command recommendations for software applications. In *Proceedings of the 22nd annual ACM symposium on User interface software and technology*. 193–202.

[71] Daniel C McFarlane and Kara A Latorella. 2002. The scope and importance of human interruption in human-computer interaction design. *Human-Computer Interaction* 17, 1 (2002), 1–61.

[72] Ananya Gubbi Mohanbabu, Rosiana Natalie, Brandon Kim, Anhong Guo, and Amy Pavel. 2026. A11y-CUA Dataset: Characterizing the Accessibility Gap in Computer Use Agents. *arXiv preprint arXiv:2602.09310* (2026).

[73] Jakob Nielsen and Rolf Molich. 1990. Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 249–256.

[74] Tianyue Ou, Wanyao Guo, Apurva Gandhi, Graham Neubig, and Xiang Yue. 2025. AgentDiagnose: An Open Toolkit for Diagnosing LLM Agent Trajectories. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Ivan Habernal, Peter Schulam, and Jörg Tiedemann (Eds.). Association for Computational Linguistics, Suzhou, China, 207–215. https://doi.org/10.18653/v1/2025.emnlp-demos.15

[75] Yi-Hao Peng, Dingzeyu Li, Jeffrey P Bigham, and Amy Pavel. 2025. Morae: Proactively Pausing UI Agents for User Choices. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25)*. Association for Computing Machinery, New York, NY, USA, Article 198, 14 pages. https://doi.org/10.1145/3746059.3747797

[76] Zachary Pousman and John Stasko. 2006. A taxonomy of ambient information systems: four patterns of design. In *Proceedings of the Working Conference on Advanced Visual Interfaces* (Venezia, Italy) *(AVI '06)*. Association for Computing Machinery, New York, NY, USA, 67–74. https://doi.org/10.1145/1133265.1133277

[77] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[78] Joshua S Rubinstein, David E Meyer, and Jeffrey E Evans. 2001. Executive control of cognitive processes in task switching. *Journal of experimental psychology: human perception and performance* 27, 4 (2001), 763.

[79] Pascal J Sager, Benjamin Meyer, Peng Yan, Rebekka von Wartburg-Kottler, Layan Etaiwi, Aref Enayati, Gabriel Nobel, Ahmed Abdulkadir, Benjamin F Grewe, and Thilo Stadelmann. 2026. A comprehensive survey of agents for computer use: Foundations, challenges, and future directions. *Journal of Artificial Intelligence Research* 85 (2026).

[80] Nitin Sawhney and Chris Schmandt. 2000. Nomadic radio: speech and audio interaction for contextual messaging in nomadic environments. *ACM transactions on Computer-Human interaction (TOCHI)* 7, 3 (2000), 353–383.

[81] Anthony R Selkowitz, Shan G Lakhmani, and Jessie YC Chen. 2017. Using agent transparency to support situation awareness of the Autonomous Squad Member. *Cognitive Systems Research* 46 (2017), 13–25.

[82] Jacob Somervell, CM Chewar, and D Scott McCrickard. 2002. Evaluating graphical vs. textual secondary displays for information notification. In *Proceedings of the ACM Southeast Conference, Raleigh NC*. 153–160.

[83] Amanda Swearngin, Jason Wu, Xiaoyi Zhang, Esteban Gomez, Jen Coughenour, Rachel Stukenborg, Bhavya Garg, Greg Hughes, Adriana Hilliard, Jeffrey P. Bigham, and Jeffrey Nichols. 2024. Towards Automated Accessibility Report Generation for Mobile Apps. *ACM Trans. Comput.-Hum. Interact.* 31, 4, Article 54 (Sept. 2024), 44 pages. https://doi.org/10.1145/3674967

[84] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, textual or hybrid: the effect of user expertise on different explanations. In *Proceedings of the 26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 109–119. https://doi.org/10.1145/3397481.3450662

[85] Maryam Taeb, Amanda Swearngin, Eldon Schoop, Ruijia Cheng, Yue Jiang, and Jeffrey Nichols. 2024. AXNav: Replaying Accessibility Tests from Natural Language. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery,

New York, NY, USA, Article 962, 16 pages. https://doi.org/10.1145/3613904.3642777

[86] Desney S. Tan, Brian Meyers, and Mary Czerwinski. 2004. WinCuts: manipulating arbitrary window regions for more effective use of screen space. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems* (Vienna, Austria) *(CHI EA '04)*. Association for Computing Machinery, New York, NY, USA, 1525–1528. https://doi.org/10.1145/985921.986106

[87] Koen van de Merwe, Steven Mallam, and Salman Nazir. 2024. Agent transparency, situation awareness, mental workload, and operator performance: A systematic literature review. *Human Factors* 66, 1 (2024), 180–208.

[88] Michael Vössing, Niklas Kühl, Matteo Lind, and Gerhard Satzger. 2022. Designing transparency for effective human-AI collaboration. *Information Systems Frontiers* 24, 3 (2022), 877–895.

[89] Bryan Wang, Zeyu Jin, and Gautham Mysore. 2022. Record Once, Post Everywhere: Automatic Shortening of Audio Stories for Social Media. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 14, 11 pages. https://doi.org/10.1145/3526113.3545680

[90] Xu Wang, Benjamin Lafreniere, and Tovi Grossman. 2018. Leveraging community-generated videos and command logs to classify and recommend software workflows. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[91] Zijie J Wang, Chudi Zhong, Rui Xin, Takuya Takagi, Zhi Chen, Duen Horng Chau, Cynthia Rudin, and Margo Seltzer. 2022. TimberTrek: exploring and curating sparse decision trees with interactive visualization. In *2022 IEEE visualization and visual analytics (VIS)*. IEEE, 60–64.

[92] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. 2019. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 56–65.

[93] 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.

[94] Yuhao Yang, Zhen Yang, Zi-Yi Dou, Anh Nguyen, Keen You, Omar Attia, Andrew Szot, Michael Feng, Ram Ramrakhya, Alexander Toshev, et al. 2025. Ultracua: A foundation model for computer use agents with hybrid action. *arXiv preprint arXiv:2510.17790* (2025).

[95] Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. 2024. Gpt-4v (ision) is a generalist web agent, if grounded. *arXiv preprint arXiv:2401.01614* (2024).

[96] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, et al. 2023. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854* (2023).

Table 1: Participants in our study were marked as P1-P30.

ID	Age	Gender	Major/Profession	ID	Age	Gender	Major/Profession
P1	24	Female	Data Science	P16	24	Female	Computer Science
P2	25	Female	Public Health	P17	26	Female	Electrical Computer Engineering
P3	23	Female	Quantitative Finance Risk Management	P18	28	Male	Computer Science and Engineering
P4	22	Male	Computer Science and Engineering	P19	22	Female	Computer Science and Engineering
P5	25	Female	Nutritional Science	P20	19	Female	Electrical and Computer Engineering
P6	23	Male	environment and sustainability	P21	22	Male	Computer Science and Engineering
P7	26	Male	Solution Intellgence Analyst	P22	26	Male	Electrical and Computer Engineering
P8	32	Male	Postdoc	P23	19	Male	Computer Science and Engineering
P9	24	Female	Geospatial data science	P24	24	Male	Computer Science and Engineering
P10	23	Female	Quantitative Finance and Risk Management	P25	26	Female	Industrial Engineering
P11	23	Male	Computer Science and Engineering	P26	25	Male	Electrical and Computer Engineering
P12	25	Female	Computer Science and Engineering	P27	23	Male	Computer Science and Engineering
P13	20	Female	Anthropology and astronomy	P28	18	Male	Computer Science and Engineering
P14	22	Female	Computer Science and Engineering	P29	23	Female	Computer Science and Engineering
P15	28	Female	Information Science	P30	23	Female	Accountant