Viago: Exploring Visual-Audio Modality Transitions for Social Media Consumption on the Go

Ruei-Che Chang*
Reality Labs Research, Meta
Toronto, Ontario, Canada
rueiche@umich.edu

Tovi Grossman
University of Toronto
Toronto, Ontario, Canada
tovi@dgp.toronto.edu

Carine Rognon
Reality Labs Research, Meta
Redmond, Washington, USA
carinerognon@meta.com

Michael Glueck
Reality Labs Research, Meta
Toronto, Ontario, Canada
mglueck@meta.com

Christopher Collins
Reality Labs Research, Meta
Toronto, Ontario, Canada
chriscollins@meta.com

Amy Karlson
Reality Labs Research, Meta
Redmond, Washington, USA
akkarlson@meta.com

Hemant Bhaskar Surale
Reality Labs Research, Meta
Toronto, Ontario, Canada
hemantsurale@meta.com

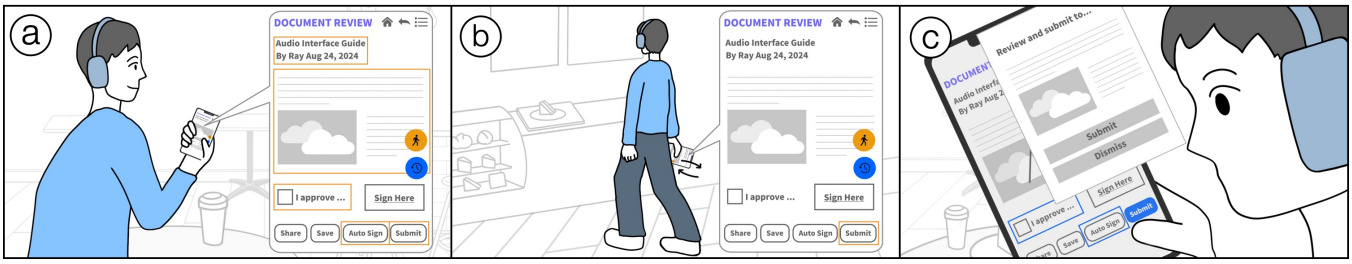

Figure 1: Viago supports visual-audio modality transitions for mobile users on the go. In a scenario where a user is waiting for an order and browsing a mobile app: (a) The user initiates Viago by tapping the orange button, which provides a visual and audio preview, preparing the user for audio interactions on the go. (b) By vaguely recalling the layout of the interface and the available audio elements, the user can navigate the screen with touch gestures and audio feedback while heading to pick up the order. The important action (*e.g.,* submit) will be cached for further visual confirmation before issuing. (c) Upon returning to the seat and resuming visual use, Viago provides visual and audio review to refresh the user's memory about content they heard and actions they did, and automatically restores cached actions, allowing the user to complete them visually.

Abstract

As mobile phone use while walking becomes increasingly prevalent, users often divide their visual attention between their surroundings and phone displays, raising concerns around safety and interaction efficiency. Alternative input and output modalities—such as eyes-free touch gestures and audio feedback—offer a promising avenue for reducing visual demands in these contexts. However, the design of seamless transitions between visual and audio modalities for mobile interaction on the go remains underexplored. To fill this gap, we conducted a design probe study with ten participants, simulating screen reader-like experiences across diverse applications to identify five key design insights and three design guidelines. Informed by these insights, we developed Viago, a background service that facilitates fluid transitions between visual and audio modalities for mobile task management while walking. A subsequent evaluation with thirteen participants demonstrated that Viago effectively supports on-the-go interactions by enabling users to interleave modalities as needed. We conclude by discussing the broader implications of visual-audio modality transitions and their potential to enhance mobile interactions in everyday, dynamic environments.

CCS Concepts

• **Human-centered computing** → **Auditory feedback**; **Gestural input**; *Accessibility technologies.*

Keywords

task interruption and resumption, mobile computing, interaction on the go, multimodal interaction, audio interface, gesture, micro-productivity, accessibility

*This work was done while the author was a Research Intern at Meta. The author is also affiliated with the Department of Computer Science and Engineering at the University of Michigan.

UIST '25, Busan, Republic of Korea

ACM ISBN 979-8-4007-2037-6/25/09
https://doi.org/10.1145/3746059.3747761

ACM Reference Format:
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 *The 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25), September 28–October 01, 2025, Busan, Republic of Korea.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3746059.3747761

1 Introduction

As mobile phone use becomes increasingly common, walking—a frequent situational impairment in everyday life [118, 138, 139]—often divides visual attention between perceiving immediate surroundings and engaging with foreground tasks, such as using mobile applications [52, 67]. This division of attention also competes for cognitive resources necessary for comprehending app content [67, 117] and impairs the fine motor skills required for precise touch inputs [64]. These factors not only present potential safety risks but also significantly hinder the ability to interact effectively with mobile apps [52, 117].

Leveraging eyes-free touch gestures and audio channels to navigate information could offer a usable and less intrusive alternative in such scenarios. Previous work indicated that the audio modality is safer, improves walking speed, and offers decent comprehension [128], while the visual modality provides more focused experiences in stable surroundings (*e.g.,* standing, sitting) [58]. The strengths of visual and audio modalities can thus be leveraged in mobile apps for on-the-go contexts. Combining these modality strengths through adaptive interfaces—such as automatic switching between visual and audio modes—has been suggested as particularly beneficial for mobile use [64, 103].

However, the feasibility and effectiveness of modality transitions, specifically moving from visually rich interfaces to lower-bandwidth audio interactions and vice versa, remain underexplored. This paper addresses this gap by focusing on three primary research questions:

(1) How can visual-to-audio transitions be designed to support interaction continuity and reduce cognitive load?
(2) How can audio-to-visual transitions be designed to support quick resumption and decision making?
(3) How effectively do users interleave visual and audio modalities when navigating a complex visual display?

To explore these questions, we started with a design probe study with ten mobile users to explore how to support the visual-audio modality transitions better by exploring different commonly-used apps. Our design probe employed visual engagement detection to automatically switch between visual mode[1] and audio mode[2]. The audio interaction model was inspired by screen reader usage, given its usability and efficacy in supporting mobile interactions through audio and gesture inputs. From this study, we derived five key design insights: support imaging visual layout and previewing audio content, provide autonomy for adding or removing audio elements to avoid audio clutter, provide hierarchical audio information and ways for skimming, provide cursors of where users left off from another mode, and provide ways to review on-the-go actions to avoid unintended ones. We organized them into three design guidelines for stages: *visual-to-audio transition* **(V2A mode)**, *audio consumption on the go* **(audio mode)** and *audio-to-visual transition* **(A2V mode)**.

Based on these insights and guidelines, we developed Viago, a smartphone application that implements interaction techniques supporting seamless visual-audio modality transitions. In V2A mode, users receive visual and audio previews, clearly highlighting elements transferable to audio mode, with options to quickly add or remove elements. In audio mode, users navigate through hierarchical audio structures using intuitive gestures, spatial audio cues, and distinct voice feedback to differentiate content types (e.g., text versus images). Before resuming normal visual interaction (A2V mode), users can efficiently review audio-mode interactions visually, complemented by previously heard audio headlines, and finalize cached actions requiring visual confirmation.

We then conducted a preliminary evaluation with thirteen participants to assess users' perceptions, usability, and the effectiveness of Viago in supporting mobile tasks on the go. They were asked to perform comprehension and selection tasks, such as identifying specific post content to *like* or *share* in a simulated social media app. The tasks were informed by the design probe study, where social media apps were common choices (Table 1), and they typically encompassed complex layouts with rich multimodal information.

We found that participants effectively managed the tasks with Viago by interleaving visual and audio modes on the go. They found the visual and audio feedback in V2A mode for preparing and adapting to audio interactions and appreciated having the autonomy to select elements of interest, which helped reduce information overload. However, many participants indicated that automatic modality transitions would be preferable over manually initiating the transition each time. For A2V mode, while users valued the ability to visually review their audio interactions, they expressed a preference for a more streamlined review method rather than scrolling through multiple posts. Participants expressed interest in the potential use of Viago with other applications (e.g., music, messaging, navigation), though they highlighted challenges such as ambient noise interfering with audio feedback and sudden interruptions disrupting seamless transitions. Based on these findings, we discuss design implications for adapting Viago's interaction techniques to additional apps and more diverse real-world scenarios.

In summary, our work contributes:

(1) Design insights and guidelines derived from a design probe study with ten users for reducing cognitive load and supporting interaction continuity and quick resumption, laying the groundwork for modality switching in on-the-go contexts.
(2) The prototype implementation of Viago, a smartphone background service that provides multimodal feedback and interaction techniques to facilitate seamless transitions between visual and audio modalities.
(3) Results from a preliminary user evaluation with thirteen mobile users suggested the limitations, feasibility, and promise of Viago.
(4) Broader design implications for future work in designing intelligent seamless modality-switching techniques in on-the-go contexts.

[1]Visual mode refers to typical visual smartphone interactions.
[2]Audio mode involves holding the phone at the user's side and primarily relying on audio feedback and touch gestures, with occasional visual glances.

2 Related Work

Our work builds upon prior works in interactions on the go with visual and audio modalities and supports for task interruption and resumption. We describe our motivation and insights drawn from prior literature below.

2.1 Interacting with Visual Information on Different Devices While Walking

Walking is a common situational impairment that often affects the use of mobile devices [62, 118, 138, 139], which divides users' attention and leads to reduced performance in both walking [76, 86, 116] and mobile tasks [12, 33, 76, 77, 79, 86, 89, 116]. To address this, prior work has explored enabling mobile interactions while walking by using adaptive information or alternative modalities. For instance, visual information on the smartphones can adapt to walking, such as by enlarging interface elements [62, 64, 75, 117, 146], customizing font configurations (*e.g.,* color, spacing, weight) [67], or adjusting the granularity of information (*e.g.,* summaries, outlines) [64, 67]. However, walking while looking down at a smartphone screen still poses potential safety risks.

In recent years, a variety of devices with different form factors have been developed to support eyes-free and hands-free information consumption and navigation. For example, head-wearable devices like smart glasses and Optical Head-Mounted Displays (OHMDs) allow users to access information while maintaining a "heads-up" posture, preserving real-world awareness. This advancement has introduced a new paradigm for on-the-go interactions, termed "heads-up computing" [156].

Prior research has explored various design dimensions for heads-up see-through displays, including adaptive visual layouts [9, 57, 96, 110, 159], distinct visual styles (*e.g.,* color, spacing, size) [39, 40, 104, 158], interface placements within the user's field of view [23, 66, 68, 96, 110], and visual notifications (*e.g.,* pictograms, text) [59, 78]. OHMDs also support text creation and editing through speech input, enabling users to avoid a "heads-down" posture while typing [19, 42, 43, 57, 159]. Other head-wearable devices, such as headphones, facilitate non-visual interactions using physical buttons, and their sensing capabilities (*e.g.,* head motion, mid-air gestures, on-surface touch) could enable rich interactions [98, 123].

Smartwatches, emerging as powerful smartphone companions [28–30], enable complex information navigation through their unique form factor [91–93, 105, 121, 140, 143]. They also support eyes-free interactions via hand motions [7, 44, 45, 145], voice input [6, 111, 157], and non-visual touch inputs [80]. Smart rings, while offering always-available and subtle input [8, 11, 41, 65, 90, 100, 151], are limited by their smaller form factor and constrained output options, relying mainly on haptic feedback [49, 60].

Despite advancements in wearable devices, they often fall short of smartphones in input and output capabilities. For example, OHMDs can be difficult to operate due to unfamiliar mid-air interactions and may continue to compromise safety while walking [22]. Smaller devices, such as smartwatches and rings, provide limited visual output and rely on alternative modalities, like audio or haptic feedback. While voice input mitigates input constraints, it raises privacy and social acceptability concerns [20, 112]. These challenges inform our decision to focus on smartphones, which remain widely used and offer socially acceptable input methods (*e.g.,* discreet gestures [34, 112, 113, 155]) and versatile output channels (*e.g.,* visual displays, haptic and audio feedback [64, 129]) to support app navigation with multimodal feedback.

2.2 Shifting Visual Information to Audio Forms

The Multiple Resources Theory (MRT) [99] and the Resource Competition Framework (RCF) [97] suggest that human cognitive resources are divided into distinct pools in mobile contexts, each with a limited capacity, typically lasting only 4 to 8 seconds. This fragmentation can lead to attention depletion and poor task performance where attentional demands overlap within the same pool (e.g., two visual tasks) but less so across pools (e.g., visual and auditory tasks). Therefore, a carefully crafted multimodal approach is essential to help users effectively manage attentional resources in mobile contexts.

While the visual channel effectively conveys complex, detailed, and spatially organized information, the linear nature of audio limits its capacity to present multiple pieces of information simultaneously. However, audio offers the distinct advantage of enabling multitasking, allowing users to listen while engaging in visual tasks [129, 147]. Prior research has explored transforming visual information into audio to free the visual sense for critical tasks. Examples include audiobooks, which convert text into audio for on-the-go consumption, and AI tools like NotebookLM [2], which transform documents into conversational podcasts or interactive queries, reducing visual strain and mitigating information overload [94]. Descriptions of the visual world are particularly beneficial for blind or visually impaired (BVI) individuals [25]. Similarly, object attributes such as distance, location, color, and size can be mapped to sound characteristics (e.g., amplitude, pitch) and delivered as music to enhance spatial awareness [31, 55].

Although audio interactions support multitasking, excessive audio content can quickly overwhelm users. Prior tools such as ROPE and Rescribe [101, 134] addressed this issue by trimming audio content into manageable segments, while other approaches harmonized overlapping audio streams (e.g., speech, notifications, music) to minimize intrusiveness [132, 133]. Chang et al.[24] explored techniques such as using distinct voices or spatial audio cues to enhance users' awareness and differentiation of mixed-reality sounds [32, 51]. Building upon these promising techniques, our work investigates shifting visual digital content into audio formats, facilitating more effective consumption during on-the-go scenarios.

2.3 Interleaving Visual and Audio Modalities for Information Navigation on the go

As mentioned, audio content consumption allows users to access and comprehend digital information from mobile devices without requiring visual attention. Such a passive eyes-free method is particularly suited to "nomadic" situations, where unobtrusive access to information is needed without disrupting the user's primary activity [114, 115]. Previous studies have explored using auditory channels to consume information while walking, showing that it is safer, improves walking speed, and offers comparable comprehension to visual reading [128]. However, while audio is effective for passive consumption, navigating complex user interfaces (UIs)

through audio feedback alone can be time-consuming and inefficient [16, 21, 64, 104, 106, 149].

This challenge is common for blind or visually impaired (BVI) individuals, who rely on screen readers for audio feedback and touch gestures to navigate mobile UIs. Screen readers use strategies like heading-level navigation to efficiently skip tedious or irrelevant details [16]. While designed primarily for individuals with disabilities, these assistive technologies can also benefit users with temporary or situational impairments, a phenomenon known as "curb-cut effect [1]." To capitalize on this potential, researchers have developed screen reader-inspired methods that minimize visual effort and conserve cognitive resources while walking. For instance, Yang et al. [149] introduced techniques enabling semantic navigation of web content through gestures for previously-visited topics or lists, benefiting both general and BVI users [106]. Similarly, Khan et al. [64] proposed "eyes-reduced" auditory skimming techniques, allowing users to consume documents in audio while walking by jumping between paragraphs or figures, complemented by brief visual glances to interact with on-screen buttons and enlarged texts.

Interleaving audio feedback with visual output based on the mobility context could enhance information consumption in on-the-go situations, as audio is less intrusive while walking [43, 114, 115, 152], while visual output provides more focused experiences in stable environments (*e.g.,* standing, sitting) [58, 152]. However, the challenge of bridging the gap between visual interfaces with high information bandwidth and the lower bandwidth of audio experiences remains under-explored [58].

In this paper, we draw inspiration from previous work on adapting screen reader experiences to support audio consumption on the go [106, 149], and explore different sound characteristics (e.g., voice fonts, spatial audio [24, 70, 84]) that beneficial to both sighted and BVI people in different contexts (e.g., web [153] or mobile use [115]). Our aim is to leverage the strengths of both visual and audio modalities and design seamless transitions between them to support managing mobile tasks in on-the-go contexts.

2.4 Task Interruption and Resumption

Mobile devices support various tasks (*e.g.,* work, learning, communication) on the go but are frequently disrupted by interruptions from multiple factors. A study of 28 mobile users performing web search tasks [97] found that cognitive resource competition significantly constrained mobile interactions, leading to fragmented attention due to environmental distractions. These distractions often result in interruptions of tasks and the need for transitions between different visual behaviors on mobile devices, such as *glance* for brief decision-making checks, *inspect* for sustained task engagement (*e.g.,* texting, reading), or *drift* for daydreaming or pondering [58]. These visual behaviors in short bursts of time could benefit micro productivity tasks, such as writing [56, 91, 127], learning [18, 37, 58], or programming [137]. However, it could create high overhead for users to resume their original tasks if interrupted [17, 71, 88, 122]; the longer or more demanding the interruptions, the more effort it would incur for resumption [17, 35, 88].

Researchers have investigated various approaches to managing interruptions. For instance, Srivastava et al. [122] introduced reviews and previews to reduce the impact of interruptions while reading. SwitchBack [81] and EyeBookmark [61] marked where users last looked on the screen, while Phosphor [13] and Mnemonic rendering [14] highlighted changes in elements upon returning to the task. Mercury [137] supported programmers to continue their task at micro levels (*e.g.,* answering brief questions) for mobile contexts and to convert their answers back to codes when resuming their workstation.

Different from prior works exploring transitions for interruptions in single modality [61, 81, 122], different devices [137] or apps [71], we explored the design of seamless transitions between visual and audio modalities for on-the-go contexts.

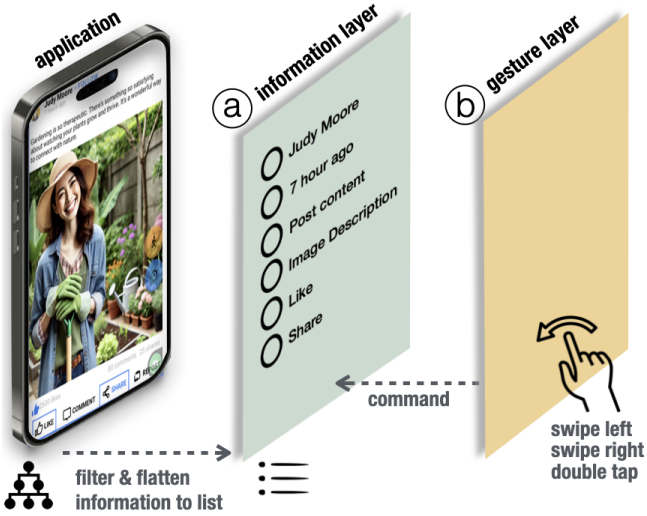

Figure 2: The audio interaction proxy used in Study 1 comprises two overlay layers: (a) The information layer parses information into a tree structure through the Android Accessibility Service, then filters out the unnecessary information (*e.g.,* layout entities) and flattens the tree nodes into a list. (b) The gesture layer allows users to navigate the information tree by swiping left or right to navigate the list elements with voice feedback and a double tap to select.

3 Study 1: Elicit Design Requirements

To explore the feasibility and design requirements for seamless visual-audio modality transitions on the go, we aimed to answer the following research question: *What design elements are critical for supporting users in transitioning tasks across modality changes?* We conducted a study with ten participants, employing a design probe that simulated screen reader-like interactions on off-the-shelf mobile apps. The probe supported automatic visual-to-audio transitions triggered by engagement detection, allowing us to elicit rich participant feedback. Below, we detail our design probe and study method.

3.1 Design Probe Supporting Audio Mode and Automatic Visual-Audio Transitions

To address our research question, we developed a design probe aimed at eliciting user insights useful in the early design phases. This probe enabled seamless switching between visual and audio interaction modes based on user engagement detection. The visual

mode mirrored typical smartphone interactions, whereas the audio mode provided a screen reader-like experience, allowing users to navigate interfaces primarily through audio feedback. In audio mode, users navigated UI elements by swiping left or right and selected items by double-tapping, guided by auditory feedback. Automatic transitions to audio mode occurred when the user's face was undetected and the phone was positioned vertically at the user's side; returning to visual mode occurred under opposite conditions. The audio mode design drew inspiration from the effective navigation practices of BVI users utilizing screen readers. The design probe consisted of two primary modules: the Audio Interaction Proxy and the Engagement Detection module.

3.1.1 Audio interaction proxy. Drawing from prior research [47, 154], we reverse-engineered an interaction proxy as an Android background service to simulate audio-based screen navigation. The proxy comprises two components: a gesture layer capturing user input and an information layer parsing current application content. The information layer continuously monitors visual content updates whenever a user input occurs through Android's Accessibility Service[3], capturing events like scrolling and button clicks. Then, it breaks down the hierarchical information of the screen into a list for gesture navigation (Figure 2a). User gestures performed on the proxy layer, including touch trajectories, finger counts, and gesture durations, generate commands to navigate the UI elements parsed by the information layer (Figure 2b). The audio interaction proxy automatically activates when users are classified as *disengaged* and deactivates upon visual re-engagement. During a pilot study with two users, we found traditional screen-reader navigation beginning from the top undesirable. Consequently, our proxy initiates audio feedback from the largest element nearest to the screen center upon activation.

3.1.2 Engagement detection. We implemented an engagement detection module to facilitate quick and automatic transitions between visual and audio interaction modes. The module classifies users as *visually engaged* when their face is detected within the front camera's field of view and the phone orientation is nearly vertical. Conversely, if no face is detected and the phone orientation shifts to nearly horizontal, the user is classified as *disengaged*. This functionality simulates features akin to Apple's *Raise to Wake* [3]. The module leverages face detection via Android's Machine Learning Kit API[4] and orientation data from the phone's inertial measurement unit (IMU). While previous work has explored various mode-switching techniques (e.g., mid-air gestures [126], touch gestures [125], multi-touch [72], pen gestures [73]), our design specifically targets automatic modality transitions suitable for on-the-go scenarios where users alternate between visually interacting with the screen and holding the device at their side while walking.

3.2 Participants

We recruited ten mobile users (6 male and 4 female), aged from 27 to 50 (Mean = 33.8). All participants had different experiences using mobile apps on the go, reported their commonly used ones, and tested them with our design probe during the study. All participants have self-reported normal auditory and motor sensory ability. The study was approved by the internal ethics board, and each participant was compensated $50 for their one-hour participation.

Table 1: Participant demographics information and the apps used in Study 1.

PID	Age	Gender	Apps used in the study
P1	27	Female	Twitter, Doordash, Notepad
P2	32	Male	Instagram, LinkedIn, Spotify
P3	28	Male	Facebook, Outlook, Spotify
P4	37	Male	Facebook, Message, Spotify
P5	32	Female	LinkedIn, Message, Spotify
P6	50	Female	Facebook, Message, LinkedIn
P7	39	Male	Facebook, Gmail, Spotify
P8	22	Male	Instagram, BBC News, Spotify
P9	38	Female	LinkedIn, Gmail, Spotify
P10	33	Male	Facebook, BBC News, Youtube

3.3 Tasks, Procedure, and Analysis

The study was conducted in person at a simulated coffee shop in our corporation. After signing informed consent forms and agreeing to be video recorded, participants described their everyday mobile usage and challenges in on-the-go scenarios and provided three commonly used apps (Table 1). They were then introduced to the design probe with their selected apps.

Participants were asked to perform their routine actions on their selected apps as mobile tasks (*e.g.*, browsing, playing music, etc.) in a three-step scenario (illustrated in Figure 1): *(i)* sitting and using their mobile phone visually while waiting for an order (30 seconds), *(ii)* walking to pick up their order while using the phone in audio mode (60 seconds), and *(iii)* returning to their seat to resume tasks in visual mode (30 seconds). Chimes indicated transitions, with participants unaware of the timing to simulate real-world experiences.

After completing the scenario, participants reflected on their experience using the design probe for mobile tasks in both visual and audio modes, brainstorming ways to improve them. This process was repeated three times for each of the three apps of their choice. Participants were encouraged to fully engage in the scenario and brainstorm creatively on *content consumption* while being informed that tasks involving *content authoring* (*e.g.*, typing) were beyond the scope of the study. The study concluded with a semi-structured exit interview that focused on the overall impressions and suggestions for using the design probe. The study took one hour for each participant.

We analyzed the recorded video footage by transcribing and coding all qualitative feedback and observations for affinity diagramming analysis.

3.4 Results - Design Insights

We discuss design feedback elicited from our participants and categorize several design insights for facilitating seamless visual-audio modality transitions. These insights go beyond traditional screen readers, focusing specifically on the transitions and their associated challenges.

[3]https://developer.android.com/reference/android/accessibilityservice/AccessibilityService
[4]https://developers.google.com/ml-kit/vision/face-detection/android

I1. Providing anchors that correspond to UI elements for visual imagery. We found that participants relied on visual imaging while on the go. For instance, several participants (P1, P2, P3, P9) would prepare themselves by reinforcing their visual memory of the screen before walking. From our post-interview, they intended to extend the visual context to the audio mode so that they could better picture which on-screen elements were being read out and where they were located, as P1 described when she got her phone up and used Twitter: *"It's nice to establish the context here where I can see. I have a visual overview of the page layout. It's hard if you have to listen to the audio and figure out they're reading the alt text for a picture."* Additionally, we observed that participants occasionally glanced at the screen while on the go to update their knowledge of the screen and locate the elements being read, as P2 described while he used Instagram on the go: *"I used a mental image to navigate through. I did a double check to ensure I knew what the screen looked like when I got up. I also kept looking at my screen to memorize the layout. So that when I scroll, I know which part of it I'm on."* In sum, having anchors to retain the visual memory of the interface when switching to audio mode is critical. It could enable users to maintain prior visual contexts and navigate their interactions on the screen without visually focusing on the smartphone.

I2. Providing hierarchical and distinct audio feedback for better navigation. The design probe allowed participants full access to the on-screen content, but this quickly became overwhelming and disorienting. All expressed the need to reduce the amount of information delivered through audio, as it was difficult to navigate. Instead, they preferred navigating through high-level information and accessing details on demand, similar to heading-level navigation of screen reader [16, 115]. For instance, P1 suggested when using Twitter on the go: *"I would prefer the author name presented first and dive into details if I am interested in this person."* Other participants shared similar thoughts, such as P9, when using LinkedIn: *"I wouldn't want all the little text that's on the screen. I just want to know the person, the title, and the option to like or repost."*, and P6 when using Facebook: *"It read out the post statistics, hashtags, and buttons but not the post title and content that I would want more. If an image, describe it to me in audio."* Moreover, participants were also confused the image descriptions versus texts in a post, as P1 stated when using Twitter *"I was unsure if it was the tweet or the image for this dog post. And I found it was image when I see it."* In sum, we should provide hierarchical information for better navigation controls that users can skim on demand, as well as distinct audio feedback for users to locate themselves rather than being drowned in all the information.

I3. Providing autonomy for determining actions on the go. Participants desired control over the switch between modes instead of an automatic switch that could be unintended. They prefer control over the automated actions aligned with past research [95, 120]. Also, besides consuming information, participants were selective in using different app functions in the audio mode. They wanted to disable the less used ones and not be read in the audio mode, as P4 remarked when using Facebook: *"Don't want the follow, recommended posts, three-dot menu, hide, view statistics to be read out. Reading out all the button labels made it confusing at first when switching to audio."* Similar comments were found across participants, such as: *"I think liking and retweeting is something I don't do so often. I'm usually very careful about what I do. I mostly use my Twitter for academic purposes."* - P1, or *"Would want to like and share the post but not comment, as I wouldn't do that while on the go."* - P6. A preference for controllability is often prioritized over automation, even if the automation accuracy is higher [107]. In sum, we should retain agency over actions that should be transferred to the audio mode, in order to align with user needs and also reduce the information overload and interaction cost.

I4. Providing visual cues and connected audio experiences for switching a focus across a modality change. Participants expressed the need for a visual indicator to resume where they left off when switching between modes. They often felt confused and disconnected when navigating the interface without it, especially unsure of what would be started with audio. For instance, when moving to audio mode, six participants suggested having a visual marker for their last visual focus or starting with high-level information they were already aware of. As P2, using LinkedIn, stated: *"I want a cursor that points out the element that I'm focused on, or that the app thinks I focused on. I'd like for it to read out the heading of that element, as soon as I put it down."* P3 also desired the ability to specify the starting point for the audio mode: *"On switching to audio, I would assume it would start reading out the top one as I typically align the one I want to read to the top. Now, it starts by the email at the center of the screen, which is not expected. It would be helpful if I can hear those feedback before switching."*

When resuming to visual mode, although some participants (P5, P8) expressed no need for feedback as they could easily reorient themselves visually, others emphasized the need for highlighting where to resume reading. As P6 mentioned: *"Maybe highlight the post for a couple of seconds to let me know which of the posts visible on the screen I was on."* Similarly to moving from visual to audio mode, P8 suggested an audio-connected experience when resuming to visual *"Same as before, remind me very briefly so that I can find my place in the app. Something as simple as the current headline you are reading is..."* In sum, we should enable users to specify what should be initiated for the audio mode first, and connect their visual and audio focus through visual cues and connected audio feedback.

I5. Providing ways to review actions taken during the audio mode. Participants tended to revisit what they had heard when returning to visual mode for several reasons. For instance, some participants chose to go back to posts that included images or videos, which were more engaging visually than audibly. P2 remarked: *"I'd go back to the panda video because it sounds funny,"* and P6 similarly noted: *"I want to revisit the image posts if I find the way it was described to me interesting."* Also, participants sought to confirm their actions while on the go. For example, P1, who used Twitter for academic purposes, stated: *"I might want visual confirmation for my actions. I don't do bookmarking often, but I think it's a useful functionality for saving and going back to look at something interesting."* P5 also felt insecure about her actions without vision and emphasized the importance of revisiting and confirming actions visually: *"I revisited what I did when I came back to visual because I wanted to understand what I did and where I was in the app. Also, I wanted to make sure I didn't accidentally like something I didn't want to."* In sum, we should provide users with ways to revisit posts they find interesting and allow them to confirm actions when they return to visual mode.

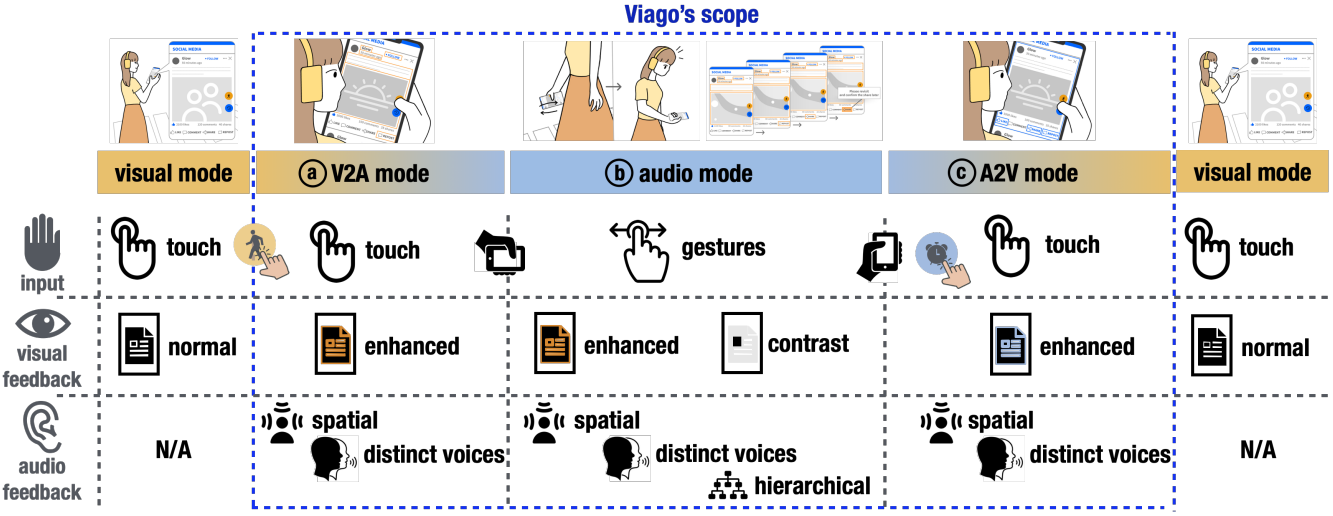

Figure 3: Overview of Viago-supported modes and their input and output features: (a) V2A mode: Activated by the orange button, providing enhanced visual and audio feedback. (b) Audio mode: Initiated when the phone is placed on the side, providing extended visual and audio feedback with hierarchical audio navigation and contrasting visuals for occasional glances. (c) A2V mode: Initiated when the user reengages with the phone visually, providing enhanced feedback for elements interacted with during the audio mode.

3.5 Design Guidelines

We identified and categorized the design insights (I1-I5) based on participants' feedback and then synthesized key design guidelines for developing interaction techniques that effectively support visual-audio modality transitions.

G1 — Visual-to-audio transition (V2A mode): Streamline rich visual screen for audio, such as by enhancing visual imagery (**I1**), determining actions of interest (**I3**), and providing visual cues and connected audio experiences to indicate where they left off (**I4**).

G2 — Content consumption on the go (Audio mode): Provide easy access to UI information, such as by providing distinct and hierarchical information (**I2**), continuing previous audio experiences (**I4**), and enhancing visual feedback for glance (**I1**).

G3 — Audio-to-visual transition (A2V mode): Streamline process of reviewing actions taken in audio, such as by enabling easy browsing of previous posts (**I5**), indicating actions taken in the audio mode (**I5**), providing visual cues and connected audio experiences to indicate where they left off in the audio mode (**I4**).

3.6 Summary

Participants expressed various needs and design feedback for visual-audio modality transitions, content consumption on the go, and preferences for certain app actions, leading to design insights and guidelines. For apps of interest, we found that social media apps (e.g., Facebook, LinkedIn) were the most common across all participants and were frequently used while on the go. Participants found using social media apps in audio mode during situational impairments acceptable due to the lower importance of social media information and the lower risk of making mistakes. In contrast, high-stakes tasks (e.g., reading or responding to a manager's email) within productivity apps (e.g., Gmail, Outlook) were preferred to be completed in a focused workspace environment rather than on the go. Also, social media apps had more complex layouts with rich text, visual media, and actionable elements than other apps with structured lists of texts (*e.g.*, messages, emails, news), while the few functions in amusement apps, such as music or video apps, were supported by headphone controls (e.g., buttons for pause, mute, or forward). These findings helped us narrow down on developing and evaluating Viago primarily on social media apps.

4 Viago Prototype

Building on the design guidelines (G1-G3), we developed Viago, a background service that enables seamless visual-audio modality transitions for consuming social media app content on the go. Below, we start with a motivating scenario of Viago and then detail the interaction techniques on visual-audio modality transitions.

4.1 Motivating Scenario

Here, we illustrate Viago in an everyday scenario, taking Karina as the main character, a professional social media influencer with multiple business collaborations.

As an influencer, Karina is highly active on social media apps and regularly partners with brands to promote their products. On her way to a meeting with her business partner in a coffee shop, she needs to complete a few unfinished tasks before their appointment. While waiting at a crosswalk, she opens her social media app and browses the posts from *Glow*, a tech brand she collaborates with and needs her publicity (Figure 4a). As the traffic light countdown reaches nine seconds, she initiates Viago by pressing the on-the-go orange button, which provides V2A transition feedback. Viago starts reading the post's headline, such as the name and timestamps, which will continue when transitioning to the audio mode (Figure 4b). The screen highlights the elements with orange bounds to

Figure 4: Viago Scenario walkthrough. (a) Karina uses her social media app with Viago running in the background while waiting at a crosswalk. (b) When the countdown reaches 9 seconds, she switches to A2V mode to consume Viago's transitioning feedback. (c) As the traffic light is about to turn green, Karina single-taps *SHARE* to transition to audio mode, skims the screen content, and confirms what is highlighted in orange by Viago for audio transfer. (d) She then puts her phone down and begins walking. (e) While walking, she can glance briefly to confirm visual content of interest with Viago enhanced visual feedback. (f) She navigates audio content using touch gestures and performs actions like clicks without needing to stare at her smartphone. (g) After reaching the bus stop, she switches to A2V mode to review actions taken in audio mode through Viago's feedback, and (h) completes high-visibility actions (*e.g.,* sharing posts). (i) Karina seamlessly alternates between visual and audio modes during her commute, adapting to different contexts.

indicate that they will be available in the audio mode, including *repost* Karina most frequently used for work and *Follow* she typically used. However, instead of frequently used ones, she needs to *share* certain posts on her other social media platforms. She quickly taps on the *share* to indicate her need for Viago to pick it up to the audio mode (Figure 4c).

At the same time, when the traffic light turns green and is about to go, Karina skims the entire screen content, especially confirming what is highlighted in orange by Viago that will be transferred to audio (Figure 4c), then puts her phone down and begins walking (Figure 4d). Viago detects disengagement, switches to audio mode, and starts reading aloud the headline of the post where Karina left off visually. Karina taps once to listen to the post content and recognizes it as one from *Glow* that she needs to repost. She navigates to the repost button by swiping right and double-taps to select it. Next, she swipes up to move to the following post from Glow, taps to hear the content, briefly glances at the accompanying image (Figure 4e), and confirms it needs to be shared on her other social media. She navigates to the share button, double-taps to share it, and receives the feedback: "Please revisit and confirm to share later. (Figure 4f)" She continues browsing other posts, checking if any require further action.

After crossing the street and reaching a bus stop, she taps the history button (blue in Figure 4g) to revisit the first post she heard in audio mode. With smooth transitioning feedback from audio to visual, Karina can easily identify the actions she performed on the go, such as the *repost*, *share*, and *like* the certain posts. When she navigates back to the post for sharing, Viago automatically prompts her to confirm and complete the sharing process (Figure 4h). Karina continues to alternate between visual and audio modes during her commute, putting her phone down when the bus becomes crowded or when the road demands more of her attention, and resuming visual mode when it is safe or convenient to do so (Figure 4i).

4.2 V2A mode: Seamless Transition from Visual to Audio Mode

To support seamless transitions from visual to audio mode (**G1**), Viago allows users to initiate interactions (**I3**) on-demand by pressing an orange button (Figure 5a). Once activated, users receive preparatory audio feedback indicating the content that will continue playing in audio mode (**I4**), visual cues identifying the UI

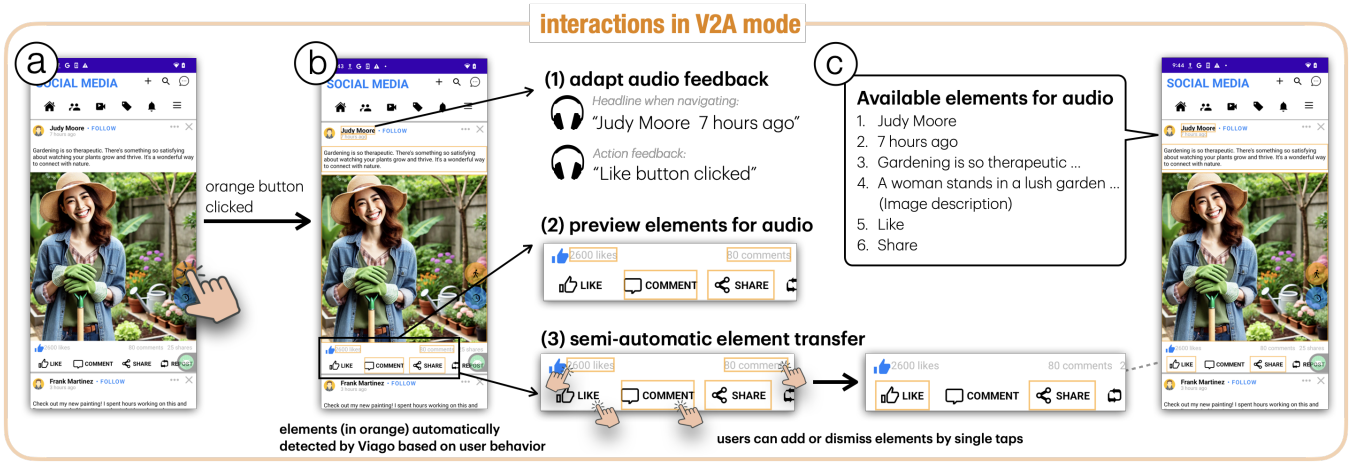

Figure 5: Interactions in V2A mode. (a) The user activates V2A mode by pressing the orange button while still visually engaged with their smartphone. (b.1) Users receive preparatory audio feedback (e.g., audio headlines and action confirmations) to facilitate a seamless transition into audio mode. (b.2) Viago visually highlights elements available for audio interaction using orange boundaries. (b.3) Users can quickly add or dismiss these highlighted elements with a single tap. (c) All elements highlighted in orange are transferred to an audio-accessible list for easy navigation on the go.

elements available for interaction in audio mode (**I1**), and the ability to selectively add or dismiss these elements with a single tap (**I3**). We describe these interaction techniques below.

4.2.1 Semi-automatic Selection of UI Buttons for Audio Mode. To support user autonomy (**I3**) in selecting elements of interest for audio interaction, Viago implements a semi-automatic mechanism based on user behavior. Specifically, it employs a time window (t) to track frequently clicked UI elements. If a particular element is clicked w times within this interval, it is automatically designated as an element to be transferred to audio mode. In our current prototype, we set $t = 1$ minute and $w = 3$. For instance, if a user clicks the LIKE button three times within one minute, Viago recognizes this action as a preference and highlights the LIKE button accordingly. After pressing the orange button to activate V2A mode, users can further customize these selections by single-tapping UI elements to add or dismiss them (Figure 5b.3). These user preferences are consistently applied across all posts.

4.2.2 Visual Feedback for Available Audio Mode Elements. After the user presses the orange button while visually engaged with the smartphone, Viago initially highlights the currently focused post with an orange boundary, indicating this will be the first item read in audio mode (**I4**), before gradually fading out. Subsequently, to enhance users' visual memory (**I1**) and provide clear previews of elements transferable to audio mode (**I4**), Viago highlights specific UI elements by default with opaque orange boundaries. These elements include the author's name, timestamp, post content, images, and other user-selected interaction buttons identified by the semi-automatic selection mechanism (Figure 5b.2).

4.2.3 Preparing Users for Audio Mode Interactions. To ensure a smooth and predictable transition into audio mode (**I4**), Viago provides preparatory audio experiences while the user is still visually interacting with content. When browsing a new post visually, the user's first exposure is accompanied by an audio headline reading the author's name and timestamp (Figure 5b.1). This same headline seamlessly continues as the initial high-level audio information when transitioning into audio mode. Additionally, users receive spatial audio feedback when interacting with UI elements; for instance, hearing "*Like button clicked*" spatially positioned on the left to reinforce visual-audio mapping and further ease cognitive load during transitions. These audio interaction principles are further detailed in the next section.

4.3 Audio Mode: Screen Navigation with Audio Feedback and Occasional Glances

Building upon V2A mode, the audio mode activates automatically when user disengagement is detected (*e.g.*, when the user's face is not detected and the phone is held at their side), as determined by the engagement detection module described in Section 3.1.2. To effectively support content consumption on the go (**G2**), Viago provides clear visual feedback optimized for quick glances (**I1**), hierarchical and distinct audio feedback (**I2**), and intuitive gestures for interaction. We detail these interaction techniques below.

4.3.1 Gestural Navigation with Hierarchical Audio Content. To facilitate eyes-free interaction while users hold the phone at their side, we adapted the audio interaction proxy (Section 3.1.1) to map simple one-handed gestures onto social media app navigation, as follows:

- **Swipe Up/Down** — Navigate to the next/previous post.
- **Swipe Right/Left** — Navigate between elements within the current focused post.
- **Single Tap** — Enter or exit the focused post.
- **Double Tap** — Select or activate a focused element.

To support efficient audio browsing (**I2**), Viago structures audio information hierarchically, allowing users to quickly skim high-level audio headlines and dive deeper into detailed content when needed. Upon swiping up or down to navigate between posts (Figure 6a), users hear audio headlines consisting of the author's name

and timestamp (*e.g., "Judy Moore, 7 hours ago"*). This headline mirrors the audio feedback provided during the earlier visual preview (Figure 5b.1), ensuring consistency and continuity across modality transitions (**I4**). Users can single-tap to explore the post in more detail, triggering Viago to read the post's textual content by default (*e.g., Starting your garden journey?..."* in Figure 6b). Within a post, users navigate among specific elements by swiping left or right (Figure 6c), and activate selected elements with a double-tap (Figure 6d). Importantly, audio navigation is limited to the UI elements explicitly chosen during the visual preview (Figure 5c). Actions with high visibility or those requiring additional confirmation (*e.g.,* sharing, reposting) are temporarily cached, awaiting user confirmation upon return to visual mode (Section 4.4).

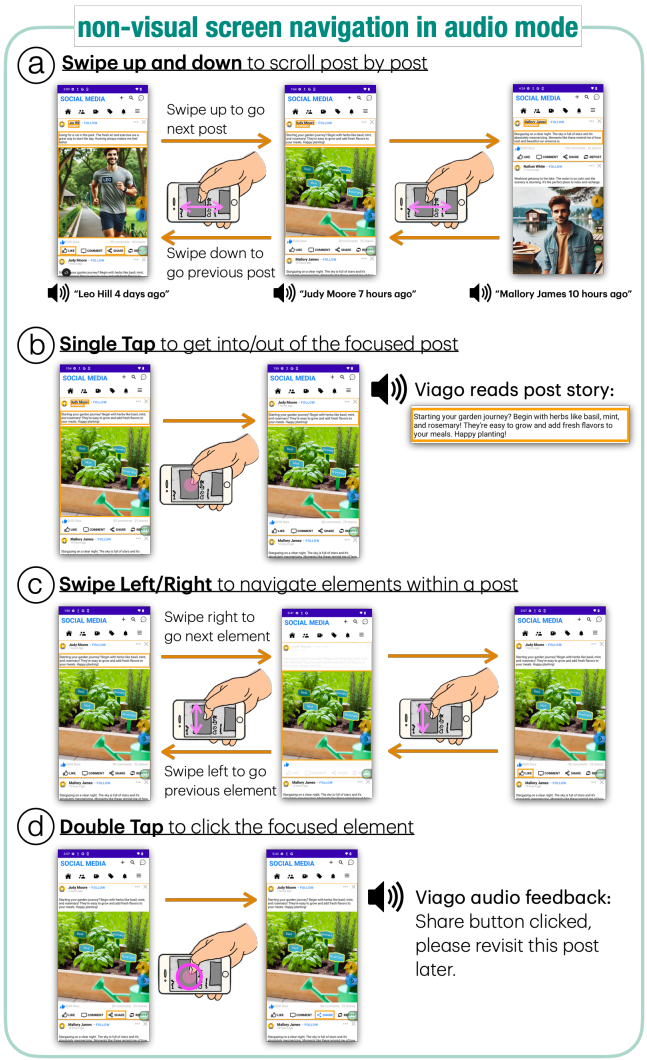

Figure 6: Non-visual navigation in audio mode. (a) Users swipe up or down to navigate between posts, with audio headlines announced for each. (b) A single tap allows users to enter the currently focused post, prompting Viago to read the detailed content aloud. (c) Within a post, users can swipe left or right to navigate granular elements, receiving detailed audio feedback for each item. (d) Users double-tap on an element to activate or select it. Gestures can be performed anywhere on the screen without targeting specific UI elements or posts directly.

4.3.2 Distinct Visual and Audio Feedback for Visual Imagery. To facilitate quick visual confirmation and reduce cognitive load during occasional glances (**I1**), Viago visually highlights the currently focused element by dimming other UI elements. Additionally, spatial audio feedback reinforces visual associations based on the element's on-screen position; for instance, the author's name *"Judy Moore"* is audibly positioned on the left, while interactions like *repost button clicked"* come from the right. To further assist users in distinguishing content types (*e.g.,* text versus images), Viago employs a female voice for text elements and a male voice for media descriptions, a design inspired by prior research in mixed-reality auditory awareness [24]. To complement auditory feedback, subtle vibration cues are provided to help users confirm their gestural inputs.

4.4 A2V mode: Seamless Transition from Audio Back to Visual Mode

To support seamless resumption from audio back to visual mode (**G3**), Viago provides efficient navigation and review functionality. In A2V mode, users can quickly jump to the first post they heard in audio mode by tapping the blue button; tapping the button again returns users directly to the latest post. As they navigate between these previously heard posts, users visually identify and review their earlier audio interactions, which are highlighted with blue boundaries (**I5**). Users are also prompted to complete previously cached actions through automatic pop-ups that preserve their interaction state. We detail these interaction techniques below.

4.4.1 Integrated Visual and Audio Feedback for Interaction Review. To facilitate recall and continuity during transition back to visual mode (**I4**), Viago provides integrated audio-visual feedback. Users hear audio headlines identical to those presented earlier in audio mode, reinforcing memory and context continuity. Visually, Viago highlights posts users previously consumed and marks UI elements they interacted with in audio mode using clear blue boundaries, enabling rapid visual confirmation.

4.4.2 Review and Completion of Cached Actions. To align with design insight **I5**—providing mechanisms for users to review interactions prior to publishing, users can efficiently revisit their actions performed during audio mode. Tapping the blue button immediately takes users back to the first post reviewed audibly (Figure 7c), enabling them to sequentially browse subsequent posts. When reviewing posts containing high-visibility actions such as *sharing* or *reposting*, Viago automatically displays any remaining steps required to complete these cached actions, allowing users to confirm and finalize them visually with minimal effort.

4.5 Implementation Details

We implemented the Viago prototype as an Android application running on a Pixel 7 Pro, utilizing Android's built-in text-to-speech API. The implementation reused both the audio interaction proxy, including gesture and information layers (Section 3.1.1), and the engagement detection module (Section 3.1.2) from our initial design probe. Additionally, to facilitate intuitive navigation between

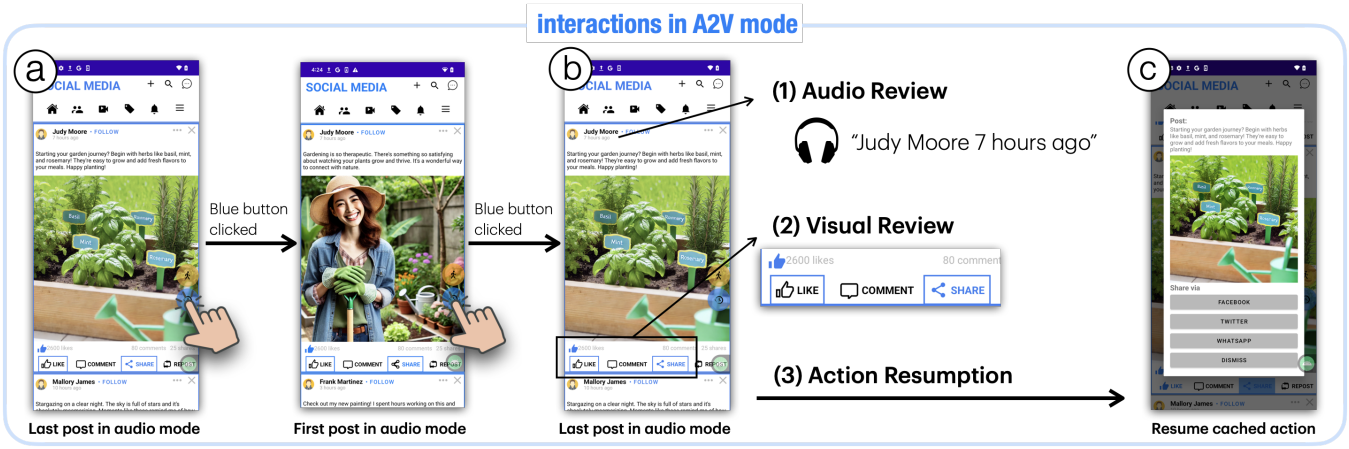

Figure 7: Interactions in A2V mode. Before fully transitioning back to visual mode: (a) Users tap the blue button to quickly jump between the first and last posts heard in audio mode, facilitating efficient visual review of previously heard posts. (b) During browsing, Viago provides audio headlines identical to those from audio mode to help users refresh their memory. Actions taken in audio mode are visually highlighted with blue boundaries. (c) For cached actions, Viago automatically presents any remaining steps, allowing users to visually confirm and complete these actions.

posts in our simulated social media app, we used the Android RecyclerView API function, `ScrollToPost`, for smooth post-by-post navigation.

5 Study 2: Preliminary Evaluation of Viago

We conducted a first-use study [50, 137] to uncover user perceptions, challenges, and design opportunities for adapting Viago to real-world use. Inspired by prior work exploring mobile work productivity given unplanned interruptions [85, 137], we simulated mobile tasks embedded in a social media app flow, incorporating "unplanned" modality-switching interruptions during tasking. Participants were prompted to interleave visual and audio modes while walking in our simulated coffee shop (same as Study 1), reflecting realistic on-the-go usage patterns. Specifically, we sought to address the following research questions:

- **RQ1:** How do participants perceive the usability of Viago for managing tasks on the go?
- **RQ2:** How does Viago support content consumption and navigation with audio feedback?
- **RQ3:** How do users perceive Viago's interactions in V2A mode?
- **RQ4:** How do users perceive Viago's interactions in A2V mode?
- **RQ5:** How could Viago's modality switching extend to other apps or real-world settings?

5.1 Participants and Apparatus

We recruited thirteen mobile users (10 male, 3 female), aged 24 to 48 (M = 35.7), with varied occupations and mobile usage experiences. All participants reported using social media apps daily and self-identified as having normal auditory and motor abilities suitable for gesture-based interaction. Each participant was provided with a Google Pixel 7 Pro running Viago and a simulated social media app, along with earphones for audio consumption. The use of a simulated app ensured task consistency across participants. The study was conducted in person at a coffee shop within our organization, with all participants visiting the location for the first time. Given the novelty of Viago, its walk-up-and-use design, and potential safety concerns in high-risk real-world environments (e.g., streets), we selected this semi-controlled setting to balance ecological validity and participant safety.

5.2 Tasks

To investigate the effectiveness of Viago, we designed tasks requiring participants to browse and comprehend content within a simulated social media app and perform specific actions under conditions of unplanned modality-switching interruptions. This approach was inspired by prior work examining mobile productivity amidst interruptions [137]. The app contained 50 posts: 12 authored by "Judy Moore," and 38 by other unique authors, each featuring distinct textual content with or without accompanying images to closely mimic real social media interactions. Participants were asked to browse and comprehend the 50 posts, and identify posts (N=12) from the author "Judy Moore" and needed to *share* her posts regarding "gardening" (N=6) and *like* her other posts not relevant to gardening (N=6) as shown in Figure 8.

During walking and browsing the app, participants periodically experienced auditory prompts (a chime) to immediately switch between visual and audio modes (the timing is denoted in Figure 8), simulating the unplanned interruptions. Actions initiated in audio mode, particularly *sharing*, required visual confirmation and completion upon returning to visual mode. Tasks were completed while participants navigated a simulated coffee shop environment (same as Study 1) with obstacles (e.g., chairs, tables, hanging lamps), replicating real-world situational impairments and divided attention. Each participant completed 12 tasks (6 shares + 6 likes) in approximately fifteen minutes.

Figure 8: Tasks for participants to like (top) or share (bottom) posts while walking in Study 2. Each dot denotes a chime to prompt participants to switch from one mode to another.

5.3 Procedure

After signing the informed consent form and receiving an overview of the study procedure, participants described their typical mobile device usage. Next, participants completed a practice session with Viago, learning to switch seamlessly between visual and audio modalities in response to audio prompts (e.g., chimes) randomly triggered by the experimenter. Participants were also instructed that they could activate the orange button at any time during visual browsing and occasionally glance at the screen in audio mode. Following this familiarization phase, participants proceeded with the primary tasks described in the previous section. The session concluded with a questionnaire and a semi-structured interview to capture participants' experiences, perceptions, and design feedback. The entire study lasted approximately 60 minutes, and each participant received compensation at a rate of $75/hour.

5.4 Data Collection

We collected qualitative data via video footage and interviews, coded through affinity diagramming, and supplemented with 7-point Likert-scale ratings to capture subjective perceptions. Also, logged task completion and success rate of *like* and *share* tasks separately, calculated as the ratio of successful instances to their total number of tasks.

5.5 Findings

We reported our findings based on each research question, and Likert-scale questions were presented in Figure 9.

5.5.1 How do participants perceive the usability of Viago to support managing mobile tasks on the go? **Participants found Viago easy to learn and use and were able to complete tasks by interleaving visual and audio mode with Viago.**

In general, participants were able to complete the tasks using Viago, including *LIKE* (Mean=91.67%) and *SHARE* posts (Mean=93.06%). Participants found Viago's visual-audio transitioning feedback intuitive, and they were able to understand and navigate the audio content on the go (Mean=6.38, SD=0.78). Some participants commented the transitioning feedback was seamless and connected across visual and audio modalities, as P2 said *"It was always on the post I expected it to be on when I switched based on the feedback. It wasn't suddenly one post lower or higher."* and P9 commended *"It's a very smooth transition in general, and I didn't find anything lagging or out of place. I also like colors for different modalities, and they don't occlude the content."* Participants also expressed that Viago was easy to learn (Mean=5.85, SD=1.07) and use (Mean=5.54, SD=1.12). However, some participants perceived the walk-up-and-use study design as somewhat challenging due to their unfamiliarity with gestures, such as *"single tap was excessive"* (P9), *"replacing single-tap with long-press by my intuition"* (P2), or *"up/down and right/left were confusing when the phone is down"* (P12). Despite this challenge, they expressed confidence in learning and adapting to Viago as time passed; P8 described his experiences in the study: *"It's unfamiliar at first because I don't really use audio interfaces. And while moving, I did not really interact with my phone. So once I got over the initial novelty, it was easy to understand."*

5.5.2 How does Viago support content consumption and navigation with audio feedback? **Participants found spatial audio feedback useful for picturing the visual layout and different voices helpful for distinguishing content sources and dividing audio content into sections while walking.**

During the study, participants tried to use audio feedback only to manage the tasks, thinking they were in the real world and had no visual access while walking. However, a few of them stopped walking or glanced on the screen a few times for different reasons, such as *"I did not want to miss a word."* - P9, *"I wanted to check the image quickly as descriptions were so long."* - P5, and *"paying attention to the audio content for where I was."* - P13, who stopped walking but still used the phone on the side. They noted the visual feedback on the screen allowed them to quickly locate the content they were listening to with a glance.

With audio feedback only, several participants (P2, P4, P6, P9) mentioned they were picturing the visual layout and thought spatial audio was helpful beyond a screen reader, as P2 noted *"The left and right audio differences were my favorites. This is futuristic to ingest posts, as opposed to a crude substitute for the visual interface. In my mental map, I was used to the name being here (pointing left) and the actions being here (pointing right). I felt almost like creating an*

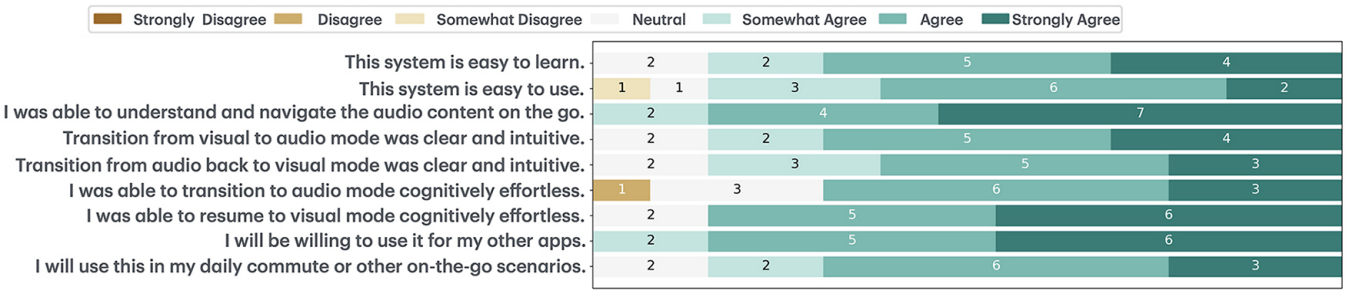

Figure 9: Likert scale questions and responses of thirteen participants in Study 2.

imaginary AR via audio rather than just using what I imagined being a screen reader user would be." Participants (P3, P6, P9, P13) also found the voice differences for text and image description useful not only for distinguishing the content type but also as a divider to help them position themselves in a post, as P3 also suggested *"It will be more like I have moved to the post section, or I am in the description section, or I am in the share section. Once you build a habit, it will be quite smooth. The brain will automatically be ready to take action to move to the next section."* Some participants also mentioned that the post content could be read with the post author's voice to enable an immersive experience for social media browsing. We will discuss more implications about audio capabilities in Section 6.1.

5.5.3 How do users perceive Viago's interactions in V2A mode? **Participants found the visual and audio preview for audio mode helpful and commended the ability to pick up elements for audio, but required smoother activation.**

Participants found the visual and audio previews in V2A mode clear (Mean=5.85, SD=1.07) and cognitively effortless (Mean=5.46, SD=1.51). Most switched to audio mode only after hearing the chime to avoid interruptions while visually browsing. However, repeatedly pressing the orange button felt cumbersome in real-world scenarios. P2, who found Viago cognitively demanding, stated: *"Because I had to put cognitive effort into thinking like I've got to click this button and then remember like what it's gonna do through the orange ones. What if this is on the street and I want to avoid something?"* In contrast, P9 and P13 pressed the orange button every time back to the visual mode, in order to prepare for unpredictable chimes, as P13 remarked: *"it tells you where you are, it tells you what's happening and knowing what you will do. I would have this mode always on if possible."* Although the short task session could not demonstrate the capability of semi-automatic element transfer that models long-term user behavior and preference, participants valued the autonomy to specify elements for audio mode, as P1 commented *"The simplicity in audio is the best. I think because you're working this linear navigation so you can't bring everything to the forefront. There's too much information there, but by having those boxes that we use most often, having controls over them, and being able to navigate while walking, it's the best part of it."*

5.5.4 How do users perceive Viago's interactions in A2V mode? **Participants found the visual review helpful and commended the ability to quickly visually review the posts they heard, but desired ways to streamline the review process.**

Participants generally found the review feedback in A2V mode clear and intuitive (Mean=5.69, SD=1.03) and cognitively effortless (Mean=6.15, SD=1.06). They found it easy to reorient after the switching due to the visual cues, as P4 remarked *"It helped when I transitioned back from audio to visual, and it had the blue boxes for me to remember where I was. The audio also reminded me of what I heard."* However, from our post-task interview, participants often overlooked audio headlines, as they focused more on the visual display during reviews.

Most participants liked the idea of visually reviewing high-visibility actions, as P13 suggested *"This is a good idea because you always have to check, you know, human nature. It is good to see what exactly you were listening to and have a review of what you've scrolled through."* Additionally, many suggested repurposing the blue button to jump to shared posts instead of starting from the first post, as not all were of interest. P7 suggested having a popup window that categorizes the actions they have done: *"Like in a notification panel, here's what you missed while you were away. You can categorize them as liked, commented, and shared posts. Those most interactive ones would be something I want to see first, and then anything else that I can just skim through."* Some participants (P2, P6, P12) also suggested adding a button to tag content of interest while walking and using distinct colors to indicate them (*e.g.,* red) other than current orange and blue cues.

5.5.5 How could Viago support visual-audio modality transitions in other apps or real-world settings? **Participants found a wider adoption of Viago to other apps and highlighted concerns and its potential for being used in real-world scenarios.**

Participants were generally willing to apply Viago to other apps (Mean=6.3, SD=0.75) and saw its potential in text-based applications like news (P4), music (P5), messaging (P2, P5, P6, P11), and navigation (P12) apps. P8 suggested it would work well for text-based social media: *"It would be very helpful with Twitter or Threads because it's text-based... less so for image-based apps like Instagram, where I prefer seeing the pictures and am less incentivized to keep it on the side."*

Though the study took place in a controlled environment, participants were also open to using Viago on the go (Mean=5.77, SD=1.01). P7 remarked: *"I could totally imagine a situation like if I'm driving or biking. It'd be super helpful. And I could just use gestures to, next, back, choose an app, hold it to get to a voice initiated."* However, some expressed concerns about real-world challenges, as P9 said *"There are uncontrollable factors out there. In here, everything is quiet.*

You're not going to bump into anything. But outside of here, are you going to accidentally bump into somebody when you're so focused on just listening to the audio?" P13 echoed this and expressed that Viago could be less useful the context requiring more cognitive focus: *" if I'm in the middle of New York City, like Times Square, in a rush hour. I'm not sure I'll be able to do this. Because I'm going to, I'm paying attention to rush hour and trying to catch a train."*

Despite these concerns, most participants still saw the potential for real-world use, drawing on their experiences in the study. For instance, P2, when avoiding a hanging lamp in the lab and putting the phone down, naturally switched to audio mode: *"When I was about bumping my head on the lanterns there, then I realized, oh, this was much more seamless in audio mode. So that gave me the kind of feeling of a real-world instinct of when I'd want to switch if there's a physical obstacle in the way."*

6 Discussion and Future Work

We discuss the limitations of our work and design implications for seamless visual-audio modality transitions.

6.1 Design Implications

Based on our study results, we propose design implications for future works that make visual-audio modality transitions generalizable to different apps and real-world contexts. This includes various components, such as automating the extraction of the UI hierarchy for audio content, designing intelligent and fluid modality transitions to minimize user effort, personalizing input methods to navigate non-visual content, developing context-effective methods to consume visual and audio output, and the potential to work across different devices.

1) Automating the creation of audio information hierarchy for facilitating consumption in different apps. Participants from Study 1 reported information overload when navigating all UI elements (**I2**), while participants from Study 2 valued the ability to skim information with audio headlines, echoed with prior findings that hierarchical audio information aids on-the-go use [115]. However, determining information hierarchy across apps is challenging due to the lack of unified norms for developers or UI metadata standards [10, 27, 148]. This implementation challenge motivated us to simulate a social media app with full UI metadata access. Previous research on audio browsing of large information structures [106, 149] found that topic- or list-based navigation reduces distraction and cognitive load while walking, compared to element-by-element UI navigation. Generating hierarchies could leverage UI semantics to group related elements, enabling summarized presentations or non-linear navigation for clarity. Building on our study results, involving users in specifying preferences before or after hierarchy generation could also enhance usability. In recent years, researchers have explored solutions like datasets [36] and algorithms for extracting or reverse-engineering UI hierarchies [102, 136, 141]. Future work could integrate these approaches to develop systems automating hierarchy generation for different apps, potentially incorporating user feedback, and leveraging large language models to structure audio-friendly information [135].

2) Designing intelligent context-aware modality transitions. Users' attention in mobility is often fragmented [97, 99], and this fragmentation depends on the surrounding environment. Participants in Study 2 (Section 5.5.5) also highlighted that uncertainties in the real world can make transitions challenging. Smartphone sensors or online navigation services can provide contextual information, such as GPS for location or Google Maps for assessing the degree of busyness (*e.g.,* traffic, crowdedness in transit), and IMU can be used for detecting user activities (*e.g.,* standing, walking, or running). This information could help systems estimate potential interruption and resumption duration, as well as determine the appropriate degree of automation and amount of audio information. For example, semi-automatic element transfers of Viago might be less practical in busy environments where users may only have 4 seconds of attention for their phone [97]. In such cases, the system should automatically deliver only the minimal viable information in audio mode rather than presenting users with multiple options to choose from. Integrating a memory and AI reasoning component could enable Viago to learn the user's decision-making patterns over time and eventually act on their behalf with confidence [119]. Also, we could leverage Viago engagement detection module to automatically turn the screen on or off for other ecological aspects, such as power consumption for sustainability or potential privacy implications [144]. Future research should consider these contextual factors and user activities to enable more context-aware transitions.

3) Personalizing gesture interactions for different apps through interaction proxies. In Study 2 (Section 5.5.1), a few participants found gestures challenging to learn and use due to their unfamiliarity, which led to incorrect gestures or accidental taps. This could be improved in several ways, such as developing personalized familiar gestures tailored to individual preferences [109, 145], utilizing specialized gestures suitable for walking contexts without touch [63, 109], or integrating voice commands to handle errors [135]. We also observed that users' daily app experiences shaped their mental models of gestures. For example, P2 expected left/right swipes to scroll posts, similar to Tinder's primary content navigation, rather than navigating between elements. By utilizing "Interaction Proxy [154]," which modifies interactions without accessing the source code, future work could allow users to personalize gestures based on app functions and their preferences. This customization could align with the audio information hierarchy by mapping gestures to information levels, such as one-to-one shortcuts (*e.g.,* a single tap for the first level, a double tap for the second) or progressive interactions (*e.g.,* a single tap to advance levels). Future exploration could also include motion-based [46, 108, 109], finger identification [48], tilt-based [46, 87], and other gestures leveraging unique smartphone form factors [150].

4) Towards context-effective synergy between visual and audio modalities. While Viago combined visual and audio feedback for digital information in V2A and A2V modes to enhance modality transitions, real-world scenarios often demand closer synergy between these modalities to handle complex situations. For instance, navigation apps provide both visual and audio directions for users on the go, but real-world complexities such as traffic or ambient sounds may influence information consumption, as mentioned in the results of our study 2. This raises key questions: What information modalities do users need in specific contexts, and how can we effectively present information across them? When considering visual contexts, gaze tracking [5, 54] could detect the user's

focus on the phone or surroundings [82], identifying visually explored content and using audio to complement it. Although Viago supports enhanced visual feedback to facilitate glancing and provide image descriptions in audio mode, further work is needed to refine the audio experience for conveying visual media beyond text, such as generating music from image descriptions [4]. Furthermore, to address the complexities of audio in augmented and virtual environments, prior research has proposed a range of strategies [83]. These include developing hardware solutions to filter out unwanted noise [26], designing algorithms to selectively extract sounds of interest [130, 131], and employing sound manipulation techniques to distinguish real-world and virtual audio [24]. Such techniques involve adjusting acoustic transparency, enhancing specific sound features [132, 133], incorporating earcons for notifications, and aligning audio with corresponding visual cues [124]. For longer app content delivered as audio, additional techniques can segment it into manageable units while preserving semantic meaning [38, 101, 134]. These segments can then be navigated in audio mode or presented in an interleaved format across visual and audio modes based on their focus. Future work could model long-term user behavior to understand modality preferences and consider environmental factors to enhance synergy between visual and audio information delivery.

5) Towards realizing eyes-free and hands-free interactions with ubiquitous devices. In this paper, we focused exclusively on smartphones due to their prevalent use and usable input and output channels compared to other commercial wearable devices (section 2.1). However, wearable devices, such as smartwatches or OHMDs, are becoming increasingly powerful with its advanced input and output capabilities to support hands-free interactions (*e.g.,* mid-air gestures [145], on-skin input [69, 142] and output [53, 74]). Some participants (P3, P6) who had smartwatches in Study 2 mentioned that using a smartwatch could be convenient for hands-free and socially acceptable interactions, especially in situations like biking, driving, or socializing [80]. While this work specifically examined transitions between visual and audio modalities on smartphones, future research should also explore other available devices, including how to distribute information across multiple modalities (*e.g.,* audio, visual, haptic) and allocating it to different devices to enhance users' awareness of both digital content and their real-world surroundings.

6.2 Limitation of the Studies

Our design probe used a screen reader-style interaction as the default audio mode to ensure full functionality through audio feedback alone. However, we recognize that sighted users' mental models and expectations for audio interaction may differ substantially from those of typical screen reader users. Indeed, some participants desired more conversational interactions with screen content, going beyond simple linear navigation. Thus, future research should investigate alternative audio interaction methods tailored explicitly to general users' expectations and situational needs, especially considering real-world interruptions (*e.g.,* ambient noise, social contexts) [24, 115].

Also, we chose a social media app as our testbed to demonstrate Viago and design tasks for Study 2, primarily due to its popularity among participants in Study 1 and its inherently complex, multimodal layout. However, mobile applications, like Reddit, might present even richer or more structurally complex content, potentially making modality transitions and content consumption more challenging in mobile contexts. Similarly, everyday mobile tasks involving content authoring (*e.g.,* text input) were not covered in the scope of our work, and thus are worth further investigation. Prior research has explored promising methods for non-visual text input [15, 44, 91], which could potentially enhance or complement our current interaction techniques for on-the-go scenarios.

Finally, Study 2 represented a preliminary first-use investigation [50, 137] conducted within a controlled setting and with a relatively small participant group. While incorporating longer tasks, diverse real-world environments, or larger sample sizes might increase the generalizability of our findings, our methodological decisions were guided by the novelty of Viago and concerns around participant safety in high-risk real-world situations. Future studies could extend this preliminary investigation by conducting a field study or deploying Viago in diverse field settings, involving varied application contexts and larger participant populations, to better capture users' ecological interactions and reflect on how Viago supports users' attentional resource management in naturally fragmented mobile interactions characterized by frequent short bursts of attention [97].

7 Conclusion

We presented Viago, a prototype smartphone application that supports seamless visual-audio modality transitions for mobile users on the go. Our design was informed by an initial design probe study with ten mobile users, which yielded several key design insights and guidelines. In Viago's V2A mode, users receive preparatory audio feedback that smoothly transitions into audio mode, preview visual cues indicating elements available in audio interactions, and selectively add or dismiss audio elements with simple taps. When transitioning back from A2V mode, Viago enables users to efficiently review interactions performed in audio mode using visual highlights and summarized audio recaps, and automatically restores cached actions for visual confirmation and completion. A subsequent evaluation with thirteen mobile users demonstrated that Viago effectively supports task management through intuitive interleaving of visual and audio modalities. Finally, we discussed broader real-world challenges and provided design implications to guide future research on visual-audio modality transitions in mobile contexts.

Acknowledgments

We thank all participants and our other colleagues at Reality Labs Research for their feedback and support throughout the project.

References

[1] 2025. Curb cut effect. https://en.wikipedia.org/wiki/Curb_cut_effect
[2] 2025. Google NotebookLM. https://notebooklm.google.com/
[3] 2025. Use Raise to Wake on your iPhone. https://support.apple.com/en-ca/108325 Accessed: 2025-03-08.
[4] Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325* (2023).

[5] Riku Arakawa, Mayank Goel, Chris Harrison, and Karan Ahuja. 2022. RGB-DGaze: Gaze Tracking on Smartphones with RGB and Depth Data. In *Proceedings of the 2022 International Conference on Multimodal Interaction* (Bengaluru, India) *(ICMI '22)*. Association for Computing Machinery, New York, NY, USA, 329–336. doi:10.1145/3536221.3556568

[6] Riku Arakawa, Jill Fain Lehman, and Mayank Goel. 2024. PrISM-Q&A: Step-Aware Voice Assistant on a Smartwatch Enabled by Multimodal Procedure Tracking and Large Language Models. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 4, Article 180 (Nov. 2024), 26 pages. doi:10.1145/3699759

[7] Riku Arakawa, Hiromu Yakura, and Mayank Goel. 2024. PrISM-Observer: Intervention Agent to Help Users Perform Everyday Procedures Sensed using a Smartwatch. 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 15, 16 pages. doi:10.1145/3654777.3676350

[8] Daniel Ashbrook, Patrick Baudisch, and Sean White. 2011. Nenya: subtle and eyes-free mobile input with a magnetically-tracked finger ring. 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, 2043–2046. doi:10.1145/1978942.1979238

[9] Yunpeng Bai, Aleksi Ikkala, Antti Oulasvirta, Shengdong Zhao, Lucia J Wang, Pengzhi Yang, and Peisen Xu. 2024. Heads-Up Multitasker: Simulating Attention Switching On Optical Head-Mounted Displays. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 79, 18 pages. doi:10.1145/3613904.3642540

[10] Mars Ballantyne, Archit Jha, Anna Jacobsen, J. Scott Hawker, and Yasmine N. El-Glaly. 2018. Study of Accessibility Guidelines of Mobile Applications. In *Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia* (Cairo, Egypt) *(MUM '18)*. Association for Computing Machinery, New York, NY, USA, 305–315. doi:10.1145/3282894.3282921

[11] Sandra Bardot, Surya Rawat, Duy Thai Nguyen, Sawyer Rempel, Huizhe Zheng, Bradley Rey, Jun Li, Kevin Fan, Da-Yuan Huang, Wei Li, and Pourang Irani. 2021. ARO: Exploring the Design of Smart-Ring Interactions for Encumbered Hands. In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction* (Toulouse & Virtual, France) *(MobileHCI '21)*. Association for Computing Machinery, New York, NY, USA, Article 12, 11 pages. doi:10.1145/3447526.3472037

[12] Leon Barnard, Ji Soo Yi, Julie A Jacko, and Andrew Sears. 2007. Capturing the effects of context on human performance in mobile computing systems. *Personal and Ubiquitous Computing* 11 (2007), 81–96.

[13] Patrick Baudisch, Desney Tan, Maxime Collomb, Dan Robbins, Ken Hinckley, Maneesh Agrawala, Shengdong Zhao, and Gonzalo Ramos. 2006. Phosphor: explaining transitions in the user interface using afterglow effects. In *Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology* (Montreux, Switzerland) *(UIST '06)*. Association for Computing Machinery, New York, NY, USA, 169–178. doi:10.1145/1166253.1166280

[14] Anastasia Bezerianos, Pierre Dragicevic, and Ravin Balakrishnan. 2006. Mnemonic rendering: an image-based approach for exposing hidden changes in dynamic displays. In *Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology* (Montreux, Switzerland) *(UIST '06)*. Association for Computing Machinery, New York, NY, USA, 159–168. doi:10.1145/1166253.1166279

[15] Syed Masum Billah, Yu-Jung Ko, Vikas Ashok, Xiaojun Bi, and IV Ramakrishnan. 2019. Accessible Gesture Typing for Non-Visual Text Entry on Smartphones. 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, 1–12. doi:10.1145/3290605.3300606

[16] Yevgen Borodin, Jeffrey P. Bigham, Glenn Dausch, and I. V. Ramakrishnan. 2010. More than meets the eye: a survey of screen-reader browsing strategies. In *Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A)* (Raleigh, North Carolina) *(W4A '10)*. Association for Computing Machinery, New York, NY, USA, Article 13, 10 pages. doi:10.1145/1805986.1806005

[17] Jelmer P. Borst, Niels A. Taatgen, and Hedderik van Rijn. 2015. What Makes Interruptions Disruptive? A Process-Model Account of the Effects of the Problem State Bottleneck on Task Interruption and Resumption. 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, 2971–2980. doi:10.1145/2702123.2702156

[18] Carrie J. Cai, Anji Ren, and Robert C. Miller. 2017. WaitSuite: Productive Use of Diverse Waiting Moments. *ACM Trans. Comput.-Hum. Interact.* 24, 1, Article 7 (mar 2017), 41 pages. doi:10.1145/3044534

[19] Runze Cai, Nuwan Janaka, Yang Chen, Lucia Wang, Shengdong Zhao, and Can Liu. 2024. PANDALens: Towards AI-Assisted In-Context Writing on OHMD During Travels. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1053, 24 pages. doi:10.1145/3613904.3642320

[20] Runze Cai, Nuwan Nanayakkarawasam Peru Kandage Janaka, Shengdong Zhao, and Minghui Sun. 2023. ParaGlassMenu: Towards Social-Friendly Subtle Interactions in Conversations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 721, 21 pages. doi:10.1145/3544548.3581065

[21] Scott Carter and Laurent Denoue. 2009. SeeReader: An (Almost) Eyes-Free Mobile Rich Document Viewer. *arXiv preprint arXiv:0909.2185* (2009).

[22] Zoe YS Chan, Aislinn JC MacPhail, Ivan PH Au, Janet H Zhang, Ben MF Lam, Reed Ferber, and Roy TH Cheung. 2019. Walking with head-mounted virtual and augmented reality devices: effects on position control and gait biomechanics. *PLoS one* 14, 12 (2019), e0225972.

[23] Chiao-Ju Chang, Yu Lun Hsu, Wei Tian Mireille Tan, Yu-Cheng Chang, Pin Chun Lu, Yu Chen, Yi-Han Wang, and Mike Y. Chen. 2024. Exploring Augmented Reality Interface Designs for Virtual Meetings in Real-world Walking Contexts. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (IT University of Copenhagen, Denmark) *(DIS '24)*. Association for Computing Machinery, New York, NY, USA, 391–408. doi:10.1145/3643834.3661538

[24] 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* (IT University of Copenhagen, Denmark) *(DIS '24)*. Association for Computing Machinery, New York, NY, USA, 116–132. doi:10.1145/3643834.3661556

[25] 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. doi:10.1145/3654777.3676375

[26] Ishan Chatterjee, Maruchi Kim, Vivek Jayaram, Shyamnath Gollakota, Ira Kemelmacher, Shwetak Patel, and Steven M Seitz. 2022. ClearBuds: wireless binaural earbuds for learning-based speech enhancement. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*. 384–396.

[27] Jieshan Chen, Chunyang Chen, Zhenchang Xing, Xiwei Xu, Liming Zhu, Guoqiang Li, and Jinshui Wang. 2020. Unblind your apps: Predicting natural-language labels for mobile gui components by deep learning. In *Proceedings of the ACM/IEEE 42nd international conference on software engineering*. 322–334.

[28] Xiang 'Anthony' Chen, Tovi Grossman, Daniel J. Wigdor, and George Fitzmaurice. 2014. Duet: exploring joint interactions on a smart phone and a smart watch. 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, 159–168. doi:10.1145/2556288.2556955

[29] Pei-Yu (Peggy) Chi and Yang Li. 2015. Weave: Scripting Cross-Device Wearable Interaction. 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, 3923–3932. doi:10.1145/2702123.2702451

[30] Pei-Yu (Peggy) Chi, Yang Li, and Björn Hartmann. 2016. Enhancing Cross-Device Interaction Scripting with Interactive Illustrations. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5482–5493. doi:10.1145/2858036.2858382

[31] Hyunsung Cho, Naveen Sendhilnathan, Michael Nebeling, Tianyi Wang, Purnima Padmanabhan, Jonathan Browder, David Lindlbauer, Tanya R. Jonker, and Kashyap Todi. 2024. SonoHaptics: An Audio-Haptic Cursor for Gaze-Based Object Selection in XR. 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 125, 19 pages. doi:10.1145/3654777.3676384

[32] Hyunsung Cho, Alexander Wang, Divya Kartik, Emily Liying Xie, Yukang Yan, and David Lindlbauer. 2024. Auptimize: Optimal Placement of Spatial Audio Cues for Extended Reality. 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 126, 14 pages. doi:10.1145/3654777.3676424

[33] James Clawson, Thad Starner, Daniel Kohlsdorf, David P. Quigley, and Scott Gilliland. 2014. Texting while walking: an evaluation of mini-qwerty text input while on-the-go. In *Proceedings of the 16th International Conference on Human-Computer Interaction with Mobile Devices & Services* (Toronto, ON, Canada) *(MobileHCI '14)*. Association for Computing Machinery, New York, NY, USA, 339–348. doi:10.1145/2628363.2628408

[34] Enrico Costanza, Samuel A. Inverso, and Rebecca Allen. 2005. Toward subtle intimate interfaces for mobile devices using an EMG controller. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Portland, Oregon, USA) *(CHI '05)*. Association for Computing Machinery, New York, NY, USA, 481–489. doi:10.1145/1054972.1055039

[35] Mary Czerwinski, Eric Horvitz, and Susan Wilhite. 2004. A diary study of task switching and interruptions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vienna, Austria) *(CHI '04)*. Association for Computing Machinery, New York, NY, USA, 175–182. doi:10.1145/985692.985715

[36] Biplab Deka, Zifeng Huang, Chad Franzen, Joshua Hibschman, Daniel Afergan, Yang Li, Jeffrey Nichols, and Ranjitha Kumar. 2017. Rico: A Mobile App Dataset for Building Data-Driven Design Applications. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (Québec City, QC, Canada) *(UIST '17)*. Association for Computing Machinery, New York, NY, USA, 845–854. doi:10.1145/3126594.3126651

[37] Tilman Dingler, Dominik Weber, Martin Pielot, Jennifer Cooper, Chung-Cheng Chang, and Niels Henze. 2017. Language learning on-the-go: opportune moments and design of mobile microlearning sessions. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Vienna, Austria) *(MobileHCI '17)*. Association for Computing Machinery, New York, NY, USA, Article 28, 12 pages. doi:10.1145/3098279.3098565

[38] Markus Frohmann, Igor Sterner, Ivan Vulić, Benjamin Minixhofer, and Markus Schedl. 2024. Segment any text: A universal approach for robust, efficient and adaptable sentence segmentation. *arXiv preprint arXiv:2406.16678* (2024).

[39] Joseph L Gabbard, J Edward Swan, and Deborah Hix. 2006. The effects of text drawing styles, background textures, and natural lighting on text legibility in outdoor augmented reality. *Presence* 15, 1 (2006), 16–32.

[40] Joseph L Gabbard, J Edward Swan, Deborah Hix, Si-Jung Kim, and Greg Fitch. 2007. Active text drawing styles for outdoor augmented reality: A user-based study and design implications. In *2007 IEEE Virtual Reality Conference*. IEEE, 35–42.

[41] Bogdan-Florin Gheran and Radu-Daniel Vatavu. 2020. From Controls on the Steering Wheel to Controls on the Finger: Using Smart Rings for In-Vehicle Interactions. In *Companion Publication of the 2020 ACM Designing Interactive Systems Conference* (Eindhoven, Netherlands) *(DIS' 20 Companion)*. Association for Computing Machinery, New York, NY, USA, 299–304. doi:10.1145/3393914.3395851

[42] Debjyoti Ghosh, Pin Sym Foong, Shengdong Zhao, Di Chen, and Morten Fjeld. 2018. EDITalk: Towards Designing Eyes-free Interactions for Mobile Word Processing. 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, 1–10. doi:10.1145/3173574.3173977

[43] Debjyoti Ghosh, Pin Sym Foong, Shengdong Zhao, Can Liu, Nuwan Janaka, and Vinitha Erusu. 2020. EYEditor: Towards On-the-Go Heads-Up Text Editing Using Voice and Manual Input. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376173

[44] Jun Gong, Zheer Xu, Qifan Guo, Teddy Seyed, Xiang 'Anthony' Chen, Xiaojun Bi, and Xing-Dong Yang. 2018. WrisText: One-handed Text Entry on Smartwatch using Wrist Gestures. 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. doi:10.1145/3173574.3173755

[45] Jun Gong, Xing-Dong Yang, and Pourang Irani. 2016. WristWhirl: One-handed Continuous Smartwatch Input using Wrist Gestures. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (Tokyo, Japan) *(UIST '16)*. Association for Computing Machinery, New York, NY, USA, 861–872. doi:10.1145/2984511.2984563

[46] João Guerreiro, Dragan Ahmetovic, Kris M. Kitani, and Chieko Asakawa. 2017. Virtual Navigation for Blind People: Building Sequential Representations of the Real-World. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility* (Baltimore, Maryland, USA) *(ASSETS '17)*. Association for Computing Machinery, New York, NY, USA, 280–289. doi:10.1145/3132525.3132545

[47] Aakar Gupta, Muhammed Anwar, and Ravin Balakrishnan. 2016. Porous Interfaces for Small Screen Multitasking using Finger Identification. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (Tokyo, Japan) *(UIST '16)*. Association for Computing Machinery, New York, NY, USA, 145–156. doi:10.1145/2984511.2984557

[48] Aakar Gupta, Muhammed Anwar, and Ravin Balakrishnan. 2016. Porous interfaces for small screen multitasking using finger identification. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. 145–156.

[49] Teng Han, Qian Han, Michelle Annett, Fraser Anderson, Da-Yuan Huang, and Xing-Dong Yang. 2017. Frictio: Passive Kinesthetic Force Feedback for Smart Ring Output. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (Québec City, QC, Canada) *(UIST '17)*. Association for Computing Machinery, New York, NY, USA, 131–142. doi:10.1145/3126594.3126622

[50] Björn Hartmann, Leith Abdulla, Manas Mittal, and Scott R. Klemmer. 2007. Authoring sensor-based interactions by demonstration with direct manipulation and pattern recognition. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. Association for Computing Machinery, New York, NY, USA, 145–154. doi:10.1145/1240624.1240646

[51] Florian Heller and Johannes Schöning. 2018. NavigaTone: Seamlessly Embedding Navigation Cues in Mobile Music Listening. 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, 1–7. doi:10.1145/3173574.3174211

[52] Juan David Hincapié-Ramos and Pourang Irani. 2013. CrashAlert: enhancing peripheral alertness for eyes-busy mobile interaction while walking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 3385–3388. doi:10.1145/2470654.2466463

[53] Da-Yuan Huang, Ruizhen Guo, Jun Gong, Jingxian Wang, John Graham, De-Nian Yang, and Xing-Dong Yang. 2017. RetroShape: Leveraging Rear-Surface Shape Displays for 2.5D Interaction on Smartwatches. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (Québec City, QC, Canada) *(UIST '17)*. Association for Computing Machinery, New York, NY, USA, 539–551. doi:10.1145/3126594.3126610

[54] Michael Xuelin Huang, Jiajia Li, Grace Ngai, and Hong Va Leong. 2017. ScreenGlint: Practical, In-situ Gaze Estimation on Smartphones. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 2546–2557. doi:10.1145/3025453.3025794

[55] Kohki Ikeuchi, Mohammed AlSada, and Tatsuo Nakajima. 2015. Providing ambient information as comfortable sound for reducing cognitive overload. In *Proceedings of the 12th International Conference on Advances in Computer Entertainment Technology*. 1–5.

[56] Shamsi T. Iqbal, Jaime Teevan, Dan Liebling, and Anne Loomis Thompson. 2018. Multitasking with Play Write, a Mobile Microproductivity Writing Tool. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (Berlin, Germany) *(UIST '18)*. Association for Computing Machinery, New York, NY, USA, 411–422. doi:10.1145/3242587.3242611

[57] Nuwan Janaka, Jie Gao, Lin Zhu, Shengdong Zhao, Lan Lyu, Peisen Xu, Maximilian Nabokow, Silang Wang, and Yanch Ong. 2023. GlassMessaging: Towards Ubiquitous Messaging Using OHMDs. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 3 (2023), 1–32.

[58] Nuwan Janaka, Xinke Wu, Shan Zhang, Shengdong Zhao, and Petr Slovak. 2022. Visual Behaviors and Mobile Information Acquisition. arXiv:2202.02748 [cs.HC] https://arxiv.org/abs/2202.02748

[59] Nuwan Nanayakkarawasam Peru Kandage Janaka, Shengdong Zhao, and Shardul Sapkota. 2023. Can Icons Outperform Text? Understanding the Role of Pictograms in OHMD Notifications. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 575, 23 pages. doi:10.1145/3544548.3580891

[60] Seungwoo Je, Minkyeong Lee, Yoonji Kim, Liwei Chan, Xing-Dong Yang, and Andrea Bianchi. 2018. PokeRing: Notifications by Poking Around the Finger. 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, 1–10. doi:10.1145/3173574.3174116

[61] Jaemin Jo, Bohyoung Kim, and Jinwook Seo. 2015. EyeBookmark: Assisting Recovery from Interruption during Reading. 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, 2963–2966. doi:10.1145/2702123.2702340

[62] Shaun K. Kane, Jacob O. Wobbrock, and Ian E. Smith. 2008. Getting off the treadmill: evaluating walking user interfaces for mobile devices in public spaces. In *Proceedings of the 10th International Conference on Human Computer Interaction with Mobile Devices and Services* (Amsterdam, The Netherlands) *(MobileHCI '08)*. Association for Computing Machinery, New York, NY, USA, 109–118. doi:10.1145/1409240.1409253

[63] Peiqi Kang, Jinxuan Li, Bingfei Fan, Shuo Jiang, and Peter B Shull. 2021. Wrist-worn hand gesture recognition while walking via transfer learning. *IEEE Journal of Biomedical and Health Informatics* 26, 3 (2021), 952–961.

[64] Taslim Arefin Khan, Dongwook Yoon, and Joanna McGrenere. 2020. Designing an Eyes-Reduced Document Skimming App for Situational Impairments. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3313831.3376641

[65] Wolf Kienzle and Ken Hinckley. 2014. LightRing: always-available 2D input on any surface. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology* (Honolulu, Hawaii, USA) *(UIST '14)*. Association for Computing Machinery, New York, NY, USA, 157–160. doi:10.1145/2642918.2647376

[66] Elisa Maria Klose, Nils Adrian Mack, Jens Hegenberg, and Ludger Schmidt. 2019. Text presentation for augmented reality applications in dual-task situations. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 636–644.

[67] Junhan Kong, Tianyuan Cai, and Zoya Bylinskii. 2023. Improving Mobile Reading Experiences While Walking Through Automatic Adaptations and Prompted

Customization. In *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23 Adjunct)*. Association for Computing Machinery, New York, NY, USA, Article 19, 3 pages. doi:10.1145/3586182.3616666

[68] Wallace S. Lages and Doug A. Bowman. 2019. Walking with adaptive augmented reality workspaces: design and usage patterns. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 356–366. doi:10.1145/3301275.3302278

[69] Gierad Laput, Robert Xiao, and Chris Harrison. 2016. ViBand: High-Fidelity Bio-Acoustic Sensing Using Commodity Smartwatch Accelerometers. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (Tokyo, Japan) *(UIST '16)*. Association for Computing Machinery, New York, NY, USA, 321–333. doi:10.1145/2984511.2984582

[70] Cheuk Yin Phipson Lee, Zhuohao Zhang, Jaylin Herskovitz, JooYoung Seo, and Anhong Guo. 2022. CollabAlly: Accessible Collaboration Awareness in Document Editing. 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 596, 17 pages. doi:10.1145/3491102.3517635

[71] Luis Leiva, Matthias Böhmer, Sven Gehring, and Antonio Krüger. 2012. Back to the app: the costs of mobile application interruptions. In *Proceedings of the 14th International Conference on Human-Computer Interaction with Mobile Devices and Services* (San Francisco, California, USA) *(MobileHCI '12)*. Association for Computing Machinery, New York, NY, USA, 291–294. doi:10.1145/2371574.2371617

[72] G. Julian Lepinski, Tovi Grossman, and George Fitzmaurice. 2010. The design and evaluation of multitouch marking menus. 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, 2233–2242. doi:10.1145/1753326.1753663

[73] Yang Li, Ken Hinckley, Zhiwei Guan, and James A. Landay. 2005. Experimental analysis of mode switching techniques in pen-based user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Portland, Oregon, USA) *(CHI '05)*. Association for Computing Machinery, New York, NY, USA, 461–470. doi:10.1145/1054972.1055036

[74] Yi-Chi Liao, Yi-Ling Chen, Jo-Yu Lo, Rong-Hao Liang, Liwei Chan, and Bing-Yu Chen. 2016. EdgeVib: Effective Alphanumeric Character Output Using a Wrist-Worn Tactile Display. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (Tokyo, Japan) *(UIST '16)*. Association for Computing Machinery, New York, NY, USA, 595–601. doi:10.1145/2984511.2984522

[75] Ji Jung Lim and Cary Feria. 2012. Visual search on a mobile device while walking. In *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services*. 295–304.

[76] Min Lin, Rich Goldman, Kathleen J Price, Andrew Sears, and Julie Jacko. 2007. How do people tap when walking? An empirical investigation of nomadic data entry. *International journal of human-computer studies* 65, 9 (2007), 759–769.

[77] Min Lin, Kathleen J Price, R Goldman, Andrew Sears, and J Jacko. 2005. Tapping on the move-Fitts' law under mobile conditions. In *Proc. IRMA*, Vol. 5. 132–135.

[78] Andrés Lucero and Akos Vetek. 2014. NotifEye: using interactive glasses to deal with notifications while walking in public. In *Proceedings of the 11th Conference on Advances in Computer Entertainment Technology* (Funchal, Portugal) *(ACE '14)*. Association for Computing Machinery, New York, NY, USA, Article 17, 10 pages. doi:10.1145/2663806.2663824

[79] Bonnie MacKay, David Dearman, Kori Inkpen, and Carolyn Watters. 2005. Walk'n scroll: a comparison of software-based navigation techniques for different levels of mobility. In *Proceedings of the 7th international conference on Human computer interaction with mobile devices & services*. 183–190.

[80] Ville Mäkelä, Johannes Kleine, Maxine Hood, Florian Alt, and Albrecht Schmidt. 2021. Hidden Interaction Techniques: Concealed Information Acquisition and Texting on Smartphones and Wearables. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 248, 14 pages. doi:10.1145/3411764.3445504

[81] Alexander Mariakakis, Mayank Goel, Md Tanvir Islam Aumi, Shwetak N. Patel, and Jacob O. Wobbrock. 2015. SwitchBack: Using Focus and Saccade Tracking to Guide Users' Attention for Mobile Task Resumption. 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, 2953–2962. doi:10.1145/2702123.2702539

[82] Sven Mayer, Gierad Laput, and Chris Harrison. 2020. Enhancing Mobile Voice Assistants with WorldGaze. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–10. doi:10.1145/3313831.3376479

[83] Mark McGill, Stephen Brewster, David McGookin, and Graham Wilson. 2020. Acoustic Transparency and the Changing Soundscape of Auditory Mixed Reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3313831.3376702

[84] Daniel Mendes, Sofia Reis, João Guerreiro, and Hugo Nicolau. 2020. Collaborative Tabletops for Blind People: The Effect of Auditory Design on Workspace Awareness. *Proc. ACM Hum.-Comput. Interact.* 4, ISS, Article 197 (Nov. 2020), 19 pages. doi:10.1145/3427325

[85] André N Meyer, Laura E Barton, Gail C Murphy, Thomas Zimmermann, and Thomas Fritz. 2017. The work life of developers: Activities, switches and perceived productivity. *IEEE Transactions on Software Engineering* 43, 12 (2017), 1178–1193.

[86] Sachi Mizobuchi, Mark Chignell, and David Newton. 2005. Mobile text entry: relationship between walking speed and text input task difficulty. In *Proceedings of the 7th international conference on Human computer interaction with mobile devices & services*. 122–128.

[87] Farhani Momotaz and Syed Masum Billah. 2021. Tilt-Explore: Making Tilt Gestures Usable for Low-Vision Smartphone Users. 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, 1154–1168. doi:10.1145/3472749.3474813

[88] Christopher A Monk, J Gregory Trafton, and Deborah A Boehm-Davis. 2008. The effect of interruption duration and demand on resuming suspended goals. *Journal of experimental psychology: Applied* 14, 4 (2008), 299.

[89] Terhi Mustonen, Maria Olkkonen, and Jukka Hakkinen. 2004. Examining mobile phone text legibility while walking. In *CHI'04 extended abstracts on Human factors in computing systems*. 1243–1246.

[90] Suranga Nanayakkara, Roy Shilkrot, Kian Peen Yeo, and Pattie Maes. 2013. EyeRing: a finger-worn input device for seamless interactions with our surroundings. In *Proceedings of the 4th Augmented Human International Conference* (Stuttgart, Germany) *(AH '13)*. Association for Computing Machinery, New York, NY, USA, 13–20. doi:10.1145/2459236.2459240

[91] Michael Nebeling, Alexandra To, Anhong Guo, Adrian A. de Freitas, Jaime Teevan, Steven P. Dow, and Jeffrey P. Bigham. 2016. WearWrite: Crowd-Assisted Writing from Smartwatches. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 3834–3846. doi:10.1145/2858036.2858169

[92] Ali Neshati, Bradley Rey, Ahmed Shariff Mohommed Faleel, Sandra Bardot, Celine Latulipe, and Pourang Irani. 2021. BezelGlide: Interacting with Graphs on Smartwatches with Minimal Screen Occlusion. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 501, 13 pages. doi:10.1145/3411764.3445201

[93] Ali Neshati, Aaron Salo, Shariff Am Faleel, Ziming Li, Hai-Ning Liang, Celine Latulipe, and Pourang Irani. 2022. EdgeSelect: Smartwatch Data Interaction with Minimal Screen Occlusion. In *Proceedings of the 2022 International Conference on Multimodal Interaction* (Bengaluru, India) *(ICMI '22)*. Association for Computing Machinery, New York, NY, USA, 288–298. doi:10.1145/3536221.3556586

[94] Maria del Carmen Ocón Palma, Anna-Maria Seeger, and Armin Heinzl. 2020. Mitigating information overload in e-commerce interactions with conversational agents. In *Information Systems and Neuroscience: NeuroIS Retreat 2019*. Springer, 221–228.

[95] Changhoon Oh, Jungwoo Song, Jinhan Choi, Seonghyeon Kim, Sungwoo Lee, and Bongwon Suh. 2018. I lead, you help but only with enough details: Understanding user experience of co-creation with artificial intelligence. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.

[96] Jason Orlosky, Kiyoshi Kiyokawa, and Haruo Takemura. 2014. Managing mobile text in head mounted displays: studies on visual preference and text placement. *ACM SIGMOBILE Mobile Computing and Communications Review* 18, 2 (2014), 20–31.

[97] Antti Oulasvirta, Sakari Tamminen, Virpi Roto, and Jaana Kuorelahti. 2005. Interaction in 4-second bursts: the fragmented nature of attentional resources in mobile HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Portland, Oregon, USA) *(CHI '05)*. Association for Computing Machinery, New York, NY, USA, 919–928. doi:10.1145/1054972.1055101

[98] Payod Panda, Molly Jane Nicholas, David Nguyen, Eyal Ofek, Michel Pahud, Sean Rintel, Mar Gonzalez-Franco, Ken Hinckley, and Jaron Lanier. 2023. Beyond Audio: Towards a Design Space of Headphones as a Site for Interaction and Sensing. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (Pittsburgh, PA, USA) *(DIS '23)*. Association for Computing Machinery, New York, NY, USA, 904–916. doi:10.1145/3563657.3596022

[99] Raja Parasuraman and David Roy Davies. 1984. Varieties of attention. *(No Title)* (1984).

[100] Farshid Salemi Parizi, Eric Whitmire, and Shwetak Patel. 2020. AuraRing: Precise Electromagnetic Finger Tracking. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 4, Article 150 (Sept. 2020), 28 pages. doi:10.1145/3369831

[101] Amy Pavel, Gabriel Reyes, and Jeffrey P. Bigham. 2020. Rescribe: Authoring and Automatically Editing Audio Descriptions. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20)*. Association for Computing Machinery, New York, NY, USA, 747–759.

doi:10.1145/3379337.3415864

[102] Yi-Hao Peng, Peggy Chi, Anjuli Kannan, Meredith Ringel Morris, and Irfan Essa. 2023. Slide Gestalt: Automatic Structure Extraction in Slide Decks for Non-Visual Access. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 829, 14 pages. doi:10.1145/3544548.3580921

[103] Kajol Rafi. 2024. "Why do I have to look to listen?": Facilitating Accessible 'On-The-Go'Audio Streaming through Gesture Interaction and Multimodal Feedback for Users Navigating Situationally-Induced Impairments.

[104] Ashwin Ram and Shengdong Zhao. 2021. LSVP: Towards Effective On-the-go Video Learning Using Optical Head-Mounted Displays. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 1, Article 30 (mar 2021), 27 pages. doi:10.1145/3448118

[105] Bradley Rey, Kening Zhu, Simon Tangi Perrault, Sandra Bardot, Ali Neshati, and Pourang Irani. 2022. Understanding and Adapting Bezel-to-Bezel Interactions for Circular Smartwatches in Mobile and Encumbered Scenarios. *Proc. ACM Hum.-Comput. Interact.* 6, MHCI, Article 201 (Sept. 2022), 28 pages. doi:10.1145/3546736

[106] Romisa Rohani Ghahari, Mexhid Ferati, Tao Yang, and Davide Bolchini. 2012. Back navigation shortcuts for screen reader users. In *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility*. 1–8.

[107] Quentin Roy, Futian Zhang, and Daniel Vogel. 2019. Automation accuracy is good, but high controllability may be better. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–8.

[108] Jaime Ruiz and Yang Li. 2011. DoubleFlip: a motion gesture delimiter for mobile interaction. 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, 2717–2720. doi:10.1145/1978942.1979341

[109] Jaime Ruiz, Yang Li, and Edward Lank. 2011. User-defined motion gestures for mobile interaction. 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, 197–206. doi:10.1145/1978942.1978971

[110] Rufat Rzayev, Paweł W. Woźniak, Tilman Dingler, and Niels Henze. 2018. Reading on Smart Glasses: The Effect of Text Position, Presentation Type and Walking. 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, 1–9. doi:10.1145/3173574.3173619

[111] Sirat Samyoun and John Stankovic. 2022. VoiceCare: a voice-interactive cognitive assistant on a smartwatch for monitoring and assisting daily healthcare activities. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2438–2441.

[112] Shardul Sapkota, Ashwin Ram, and Shengdong Zhao. 2021. Ubiquitous Interactions for Heads-Up Computing: Understanding Users' Preferences for Subtle Interaction Techniques in Everyday Settings. In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction* (Toulouse & Virtual, France) *(MobileHCI '21)*. Association for Computing Machinery, New York, NY, USA, Article 36, 15 pages. doi:10.1145/3447526.3472035

[113] T. Scott Saponas, Chris Harrison, and Hrvoje Benko. 2011. PocketTouch: through-fabric capacitive touch input. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology* (Santa Barbara, California, USA) *(UIST '11)*. Association for Computing Machinery, New York, NY, USA, 303–308. doi:10.1145/2047196.2047235

[114] Nitin Sawhney and Chris Schmandt. 1998. Speaking and listening on the run: Design for wearable audio computing. In *Digest of Papers. Second International Symposium on Wearable Computers (Cat. No. 98EX215)*. IEEE, 108–115.

[115] 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.

[116] Siobhan M Schabrun, Wolbert van den Hoorn, Alison Moorcroft, Cameron Greenland, and Paul W Hodges. 2014. Texting and walking: strategies for postural control and implications for safety. *PloS one* 9, 1 (2014), e84312.

[117] Bastian Schildbach and Enrico Rukzio. 2010. Investigating selection and reading performance on a mobile phone while walking. In *Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services* (Lisbon, Portugal) *(MobileHCI '10)*. Association for Computing Machinery, New York, NY, USA, 93–102. doi:10.1145/1851600.1851619

[118] Andrew Sears, Mark Young, and Jinjuan Feng. 2007. Physical disabilities and computing technologies: an analysis of impairments. In *The human-computer interaction handbook*. CRC Press, 855–878.

[119] Omar Shaikh, Shardul Sapkota, Shan Rizvi, Eric Horvitz, Joon Sung Park, Diyi Yang, and Michael S Bernstein. 2025. Creating General User Models from Computer Use. *arXiv preprint arXiv:2505.10831* (2025).

[120] Ben Shneiderman. 1993. Beyond intelligent machines: just do it. *IEEE software* 10, 1 (1993), 100–103.

[121] Gaganpreet Singh, William Delamare, and Pourang Irani. 2018. D-SWIME: A Design Space for Smartwatch Interaction Techniques Supporting Mobility and Encumbrance. 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, 1–13. doi:10.1145/3173574.3174208

[122] Namrata Srivastava, Rajiv Jain, Jennifer Healey, Zoya Bylinskii, and Tilman Dingler. 2021. Mitigating the Effects of Reading Interruptions by Providing Reviews and Previews. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 229, 6 pages. doi:10.1145/3411763.3451610

[123] Dennis Stanke, Pia Brandt, and Michael Rohs. 2022. Exploring the Design Space of Headphones as Wearable Public Displays. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 295, 7 pages. doi:10.1145/3491101.3519756

[124] Xia Su, Eunyee Koh, and Chang Xiao. 2024. SonifyAR: Context-Aware Sound Effect Generation in Augmented Reality. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 297, 7 pages. doi:10.1145/3613905.3650927

[125] Hemant Bhaskar Surale, Fabrice Matulic, and Daniel Vogel. 2017. Experimental Analysis of Mode Switching Techniques in Touch-based User Interfaces. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3267–3280. doi:10.1145/3025453.3025865

[126] Hemant Bhaskar Surale, Fabrice Matulic, and Daniel Vogel. 2019. Experimental Analysis of Barehand Mid-air Mode-Switching Techniques in Virtual Reality. 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, 1–14. doi:10.1145/3290605.3300426

[127] Jaime Teevan, Shamsi T. Iqbal, and Curtis von Veh. 2016. Supporting Collaborative Writing with Microtasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2657–2668. doi:10.1145/2858036.2858108

[128] Kristin Vadas, Nirmal Patel, Kent Lyons, Thad Starner, and Julie Jacko. 2006. Reading on-the-go: a comparison of audio and hand-held displays. In *Proceedings of the 8th Conference on Human-Computer Interaction with Mobile Devices and Services* (Helsinki, Finland) *(MobileHCI '06)*. Association for Computing Machinery, New York, NY, USA, 219–226. doi:10.1145/1152215.1152262

[129] Yolanda Vazquez-Alvarez and Stephen A. Brewster. 2011. Eyes-free multitasking: the effect of cognitive load on mobile spatial audio interfaces. 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, 2173–2176. doi:10.1145/1978942.1979258

[130] Bandhav Veluri, Malek Itani, Justin Chan, Takuya Yoshioka, and Shyamnath Gollakota. 2023. Semantic Hearing: Programming Acoustic Scenes with Binaural Hearables. 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 89, 15 pages. doi:10.1145/3586183.3606779

[131] Bandhav Veluri, Malek Itani, Tuochao Chen, Takuya Yoshioka, and Shyamnath Gollakota. 2024. Look Once to Hear: Target Speech Hearing with Noisy Examples. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 37, 16 pages. doi:10.1145/3613904.3642057

[132] Alexander Wang, Yi Fei Cheng, and David Lindlbauer. 2024. MARingBA: Music-Adaptive Ringtones for Blended Audio Notification Delivery. 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 729, 15 pages. doi:10.1145/3613904.3642376

[133] Alexander Wang, David Lindlbauer, and Chris Donahue. 2024. Towards Music-Aware Virtual Assistants. 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 127, 14 pages. doi:10.1145/3654777.3676416

[134] 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. doi:10.1145/3526113.3545680

[135] Bryan Wang, Gang Li, and Yang Li. 2023. Enabling Conversational Interaction with Mobile UI using Large Language Models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 432, 17 pages. doi:10.1145/3544548.3580895

[136] Bryan Wang, Gang Li, Xin Zhou, Zhourong Chen, Tovi Grossman, and Yang Li. 2021. Screen2Words: Automatic Mobile UI Summarization with Multimodal Learning. 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, 498–510. doi:10.1145/3472749.3474765

[137] Alex C. Williams, Harmanpreet Kaur, Shamsi Iqbal, Ryen W. White, Jaime Teevan, and Adam Fourney. 2019. Mercury: Empowering Programmers' Mobile Work Practices with Microproductivity. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 81–94. doi:10.1145/3332165.3347932

[138] Jacob O. Wobbrock. 2019. Situationally aware mobile devices for overcoming situational impairments. In *Proceedings of the ACM SIGCHI Symposium on Engineering Interactive Computing Systems* (Valencia, Spain) *(EICS '19)*. Association for Computing Machinery, New York, NY, USA, Article 1, 18 pages. doi:10.1145/3319499.3330292

[139] Jacob O Wobbrock. 2019. Situationally-induced impairments and disabilities. *Web Accessibility: A Foundation for Research* (2019), 59–92.

[140] Pui Chung Wong, Kening Zhu, Xing-Dong Yang, and Hongbo Fu. 2020. Exploring Eyes-free Bezel-initiated Swipe on Round Smartwatches. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–11. doi:10.1145/3313831.3376393

[141] Jason Wu, Xiaoyi Zhang, Jeff Nichols, and Jeffrey P Bigham. 2021. Screen Parsing: Towards Reverse Engineering of UI Models from Screenshots. 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, 470–483. doi:10.1145/3472749.3474763

[142] Robert Xiao, Teng Cao, Ning Guo, Jun Zhuo, Yang Zhang, and Chris Harrison. 2018. LumiWatch: On-Arm Projected Graphics and Touch Input. 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, 1–11. doi:10.1145/3173574.3173669

[143] Robert Xiao, Gierad Laput, and Chris Harrison. 2014. Expanding the input expressivity of smartwatches with mechanical pan, twist, tilt and click. 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, 193–196. doi:10.1145/2556288.2557017

[144] Jian Xu, Syed Masum Billah, Roy Shilkrot, and Aruna Balasubramanian. 2019. DarkReader: Bridging the Gap Between Perception and Reality of Power Consumption in Smartphones for Blind Users. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility* (Pittsburgh, PA, USA) *(ASSETS '19)*. Association for Computing Machinery, New York, NY, USA, 96–104. doi:10.1145/3308561.3353806

[145] Xuhai Xu, Jun Gong, Carolina Brum, Lilian Liang, Bongsoo Suh, Shivam Kumar Gupta, Yash Agarwal, Laurence Lindsey, Runchang Kang, Behrooz Shahsavari, Tu Nguyen, Heriberto Nieto, Scott E Hudson, Charlie Maalouf, Jax Seyed Mousavi, and Gierad Laput. 2022. Enabling Hand Gesture Customization on Wrist-Worn Devices. 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 496, 19 pages. doi:10.1145/3491102.3501904

[146] Tetsuo Yamabe and Kiyotaka Takahashi. 2007. Experiments in mobile user interface adaptation for walking users. In *The 2007 International Conference on Intelligent Pervasive Computing (IPC 2007)*. IEEE, 280–284.

[147] Shingo Yamano, Takamitsu Hamajo, Shunsuke Takahashi, and Keita Higuchi. 2012. EyeSound: single-modal mobile navigation using directionally annotated music. In *Proceedings of the 3rd Augmented Human International Conference* (Megève, France) *(AH '12)*. Association for Computing Machinery, New York, NY, USA, Article 22, 4 pages. doi:10.1145/2160125.2160147

[148] Shunguo Yan and PG Ramachandran. 2019. The current status of accessibility in mobile apps. *ACM Transactions on Accessible Computing (TACCESS)* 12, 1 (2019), 1–31.

[149] Tao Yang, Mexhid Ferati, Yikun Liu, Romisa Rohani Ghahari, and Davide Bolchini. 2012. Aural browsing on-the-go: listening-based back navigation in large web architectures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 277–286. doi:10.1145/2207676.2207715

[150] Yen-Ting Yeh, Antony Albert Raj Irudayaraj, and Daniel Vogel. 2024. Single-handed Folding Interactions with a Modified Clamshell Flip Phone. 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 406, 14 pages. doi:10.1145/3613904.3642554

[151] Hui-Shyong Yeo, Juyoung Lee, Hyung-il Kim, Aakar Gupta, Andrea Bianchi, Daniel Vogel, Hideki Koike, Woontack Woo, and Aaron Quigley. 2019. WRIST: Watch-Ring Interaction and Sensing Technique for Wrist Gestures and Macro-Micro Pointing. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services* (Taipei, Taiwan) *(MobileHCI '19)*. Association for Computing Machinery, New York, NY, USA, Article 19, 15 pages. doi:10.1145/3338286.3340130

[152] Chen-Hsiang Yu and Robert C. Miller. 2011. Enhancing mobile browsing and reading. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI EA '11)*. Association for Computing Machinery, New York, NY, USA, 1783–1788. doi:10.1145/1979742.1979845

[153] Lotus Zhang, Jingyao Shao, Augustina Ao Liu, Lucy Jiang, Abigale Stangl, Adam Fourney, Meredith Ringel Morris, and Leah Findlater. 2022. Exploring Interactive Sound Design for Auditory Websites. 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 222, 16 pages. doi:10.1145/3491102.3517695

[154] Xiaoyi Zhang, Anne Spencer Ross, Anat Caspi, James Fogarty, and Jacob O. Wobbrock. 2017. Interaction Proxies for Runtime Repair and Enhancement of Mobile Application Accessibility. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 6024–6037. doi:10.1145/3025453.3025846

[155] Shengdong Zhao, Pierre Dragicevic, Mark Chignell, Ravin Balakrishnan, and Patrick Baudisch. 2007. Earpod: eyes-free menu selection using touch input and reactive audio feedback. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. Association for Computing Machinery, New York, NY, USA, 1395–1404. doi:10.1145/1240624.1240836

[156] Shengdong Zhao, Felicia Tan, and Katherine Fennedy. 2023. Heads-Up Computing Moving Beyond the Device-Centered Paradigm. *Commun. ACM* 66, 9 (aug 2023), 56–63. doi:10.1145/3571722

[157] Shiwen Zhao, Brandt Westing, Shawn Scully, Heri Nieto, Roman Holenstein, Minwoo Jeong, Krishna Sridhar, Brandon Newendorp, Mike Bastian, Sethu Raman, Tim Paek, Kevin Lynch, and Carlos Guestrin. 2019. Raise to Speak: An Accurate, Low-power Detector for Activating Voice Assistants on Smartwatches. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 2736–2744. doi:10.1145/3292500.3330761

[158] Chen Zhou, Katherine Fennedy, Felicia Fang-Yi Tan, Shengdong Zhao, and Yurui Shao. 2023. Not All Spacings are Created Equal: The Effect of Text Spacings in On-the-go Reading Using Optical See-Through Head-Mounted Displays. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 720, 19 pages. doi:10.1145/3544548.3581430

[159] Chen Zhou, Zihan Yan, Ashwin Ram, Yue Gu, Yan Xiang, Can Liu, Yun Huang, Wei Tsang Ooi, and Shengdong Zhao. 2024. GlassMail: Towards Personalised Wearable Assistant for On-the-Go Email Creation on Smart Glasses. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (IT University of Copenhagen, Denmark) *(DIS '24)*. Association for Computing Machinery, New York, NY, USA, 372–390. doi:10.1145/3643834.3660683